(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

10/534510

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
27. Mai 2004 (27.05.2004)

PCT

(10) Internationale Veröffentlichungsnummer
WO 2004/044772 A2

(51) Internationale Patentklassifikation: G06F 17/30

(21) Internationales Aktenzeichen: PCT/EP2003/011655

(22) Internationales Anmeldedatum:
21. Oktober 2003 (21.10.2003)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
102 52 445.9 12. November 2002 (12.11.2002) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **SIEMENS AKTIENGESELLSCHAFT** [DE/DE]; Wittelsbacherplatz 2, 80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: **HAFT, Michael** [DE/DE]; Grünlandstrasse 27, 85604 Zorneding (DE). **HOFMANN, Reimar** [DE/DE]; Hiltenspergerstr. 79, 80796 München (DE).

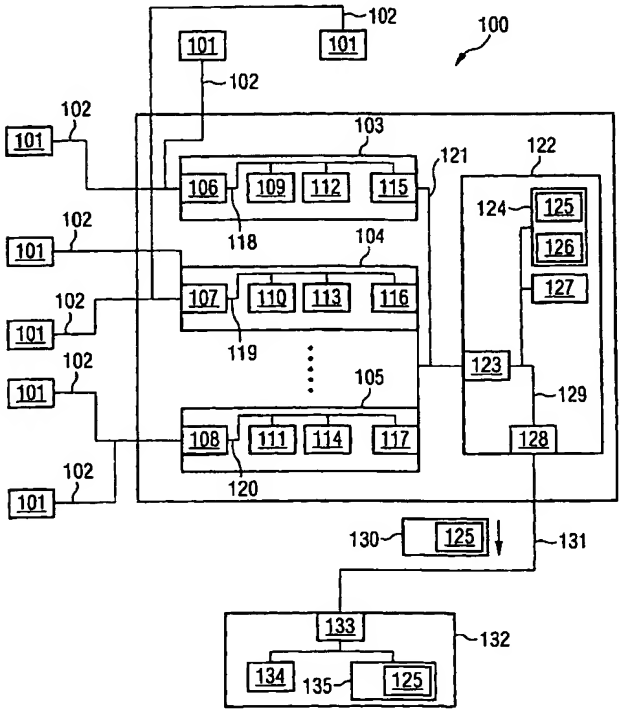
(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, 80506 München (DE).

(81) Bestimmungsstaaten *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN,

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD AND COMPUTER CONFIGURATION FOR PROVIDING DATABASE INFORMATION OF A FIRST DATABASE AND METHOD FOR CARRYING OUT THE COMPUTER-AIDED FORMATION OF A STATISTICAL IMAGE OF A DATABASE

(54) Bezeichnung: VERFAHREN UND COMPUTER-ANORDNUNG ZUM BEREITSTELLEN VON DATENBANKINFORMATION EINER ERSTEN DATENBANK UND VERFAHREN ZUM RECHNERGESTÜTZTEN BILDEN EINES STATISTISCHEN ABBILDES EINER DATENBANK



(57) Abstract: A first statistical image is formed for the first database whereby representing statistical correlations of the data elements contained in the first database. Afterwards, the first statistical image is stored in a server computer and transmitted from this server computer over a communications network to a client computer. The received first statistical image is processed by the client computer.

(57) Zusammenfassung: Für die erste Datenbank wird ein erstes statistisches Abbild gebildet, welches die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente repräsentiert. Anschließend wird das erste statistische Abbild in einem Server-Computer gespeichert und von diesem über ein Kommunikationsnetz zu einem Client-Computer übertragen. Das empfangene erste statistische Abbild wird von dem Client-Computer weiterverarbeitet.

WO 2004/044772 A2

MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) **Bestimmungsstaaten** *(regional)*: ARIPO-Patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI-Patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— *ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts*

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

Beschreibung

Verfahren und Computer-Anordnung zum Bereitstellen von Datenbankinformation einer ersten Datenbank und Verfahren zum rechnergestützten Bilden eines statistischen Abbildes einer Datenbank

Die Erfindung betrifft ein Verfahren und eine Computer-Anordnung zum Bereitstellen von Datenbankinformation einer ersten Datenbank und ein Verfahren zum rechnergestützten Bilden eines statistischen Abbildes einer Datenbank.

Heutzutage sind kaum noch Vorgänge zu beobachten, die ohne Unterstützung eines Computers ablaufen. Häufig wird bei Einsatz eines Computers im Rahmen eines Prozesses der Prozess mittels des Computers überwacht oder zumindest prozessspezifische Daten von dem Computer aufgezeichnet und protokolliert, beispielsweise Daten über die einzelnen Prozessschritte des Prozesses und deren Ergebnisse oder Zwischenergebnisse.

Beispielsweise wird üblicherweise in einem Call Center im Detail festgehalten, wann welcher Anruf in dem Call Center eingegangen ist, wann der jeweilige eingegangene Anruf von einem Mitarbeiter des Call Centers bearbeitet wurde, zu welchem anderen Mitarbeiter des Call Centers möglicherweise weitergeleitet worden ist, etc.

Ferner werden üblicherweise in der Prozess-Automatisierung umfangreiche Protokoll-Dateien gebildet, in denen Daten über die einzelnen Prozesse gespeichert werden.

Ein drittes Anwendungsgebiet ist in der Telekommunikation zu sehen; so werden beispielsweise in den Switches eines Mobilfunknetzes Protokolldaten über den in den Switches auftretenden Datenverkehr ermittelt und gespeichert.

Schließlich werden auch in einem Webserver-Computer häufig Protokolldaten über den Datenverkehr, beispielsweise über die Zugriffshäufigkeit auf von dem Webserver-Computer bereitgestellter Information, gebildet.

Treten im Verlauf eines Prozesses Probleme auf, so wird üblicherweise der Betreiber der Anlage, auf welcher der Prozess ausgeführt wird, vor Ort versuchen, die Ursache für die aufgetretenen Probleme zu finden. Gelingt ihm das nicht, so wendet er sich meist an den Hersteller der Anlage. Herstellerseitig ist es zum Auffinden der Problemursache erforderlich, auf die protokollierten Prozessdaten, allgemein auf die aufgezeichneten Protokolldaten der Anlage zuzugreifen. Derzeit hat eine die Protokolldaten enthaltende Protokolldatei eine erhebliche Größe, häufig in der Größenordnung einiger Dutzend GByte. Eine solche Protokolldatei lässt sich aus diesem Grund nur schlecht zu dem Hersteller der Anlage, beispielsweise unter Verwendung von FTP (File Transfer Protocol) übertragen. Selbst wenn ausreichend schnelle Kommunikationsverbindungen zur Verfügung stehen, ist es für den Hersteller einer Anlage schwierig und teuer, für eine größere Anzahl von Kunden die Protokolldateien zu speichern und zu verarbeiten.

Auch in anderen Bereichen besteht der Bedarf, zu Analysezwecken große Datenmengen zu übertragen, beispielsweise überall dort, wo große Datenbanken öffentlich zugänglich sind, um der Öffentlichkeit das Forschen unter Verwendung der Datenbankdaten zu ermöglichen. Die Datenbankdaten können Daten sein aus (öffentlichen) Forschungsprojekten (beispielsweise Daten einer Gen-Datenbank oder einer Protein-Datenbank), Wetterdaten, demographische Daten, Daten, die zum Zwecke einer Rasterfahndung (in diesem Fall nur einem begrenzten Kreis befugter Nutzer) zur Verfügung gestellt werden sollen. Insbesondere der Bereich der Biotechnologie ist heutzutage von erheblichem Interesse.

Es existieren eine Vielzahl von Datenbanken in diesem Bereich.

Ferner ist es insbesondere aus Gründen der Datensicherheit häufig wünschenswert, nicht alle konkreten Informationen der Datenbankdaten weiterzugeben.

Eine bekannte Möglichkeit, Informationen einer Datenbank über ein Kommunikationsnetz von einem Server-Computer einem Client-Computer bereitzustellen, besteht darin, Diagnose- oder Statistik-Werkzeuge zur Analyse der in den Datenbanken enthaltenen Daten direkt serverseitig zu installieren, welche beispielsweise unter Verwendung eines Web-Servers, welcher auf dem Server-Computer installiert ist und eines auf einem Client-Computer installierten Web-Browser-Programms genutzt werden können. Hierfür können so genannte OLAP-Werkzeuge (On-Line Analytical Processing-Werkzeuge) eingesetzt werden, deren Betrieb allerdings sehr aufwendig und teuer ist. Bei einigen OLAP-Werkzeugen ist die zu verarbeitende Datenmenge sogar schon so groß geworden, so dass die OLAP-Werkzeuge versagen.

Ferner ist es für den Betreiber einer Anlage sehr unbequem und teuer, diese Werkzeuge serverseitig zu betreiben, da das unmittelbare Interesse an der Information ja bei dem Nutzer des Client-Computers liegt und häufig der Betreiber der Anlage nicht bereit ist, die zusätzlichen Kosten für die Bereitstellung und Wartung des Server-Computers und der OLAP-Werkzeuge zu tragen.

Weiterhin ist bei einer großen Anzahl von Client-Computern und einer großen Zahl von Anfragen an den Server-Computer die Beantwortung aller Anfragen sehr rechenaufwendig, weshalb die Hardware des Server-Computers häufig unakzeptabel teuer ist.

Der Erfindung liegt das Problem eines effizienten Zugriffs auf den Inhalt einer Datenbank über ein Kommunikationsnetz

unter Wahrung der Vertraulichkeit der in der Datenbank enthaltenen Daten zugrunde.

Das Problem wird durch ein Verfahren und eine Computer-Anordnung zum Bereitstellen von Datenbankinformation einer ersten Datenbank sowie durch ein Verfahren zum rechnergestützten Bilden eines statistischen Modells einer Datenbank mit den Merkmalen gemäß den unabhängigen Patentansprüchen gelöst.

Das allgemeine Szenario, welches von der Erfindung adressiert wird, ist auf folgende Weise charakterisiert: An einem ersten Ort A steht eine große Menge von in einer Datenbank gespeicherten Daten zur Verfügung. An einem zweiten Ort B will jemand diese zur Verfügung stehenden Daten nutzen. Der Nutzer an dem Ort B ist weniger an einzelnen Datensätzen interessiert, sondern in erster Linie an der die Datenbankdaten charakterisierenden Statistik.

Bei einem Verfahren zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank wird für die erste Datenbank ein erstes statistisches Abbild beispielsweise in Form eines gemeinsamen Wahrscheinlichkeitsmodells gebildet. Dieses Abbild bzw. Modell repräsentiert die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente. Das erste statistische Abbild wird in einem Server-Computer gespeichert. Ferner wird das erste statistische Abbild von dem Server-Computer über ein Kommunikationsnetz zu einem Client-Computer übertragen und das empfangene erste statistische Abbild wird von dem Client-Computer weiterverarbeitet.

Eine Computer-Anordnung zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank weist einen Server-Computer und einen Client-Computer auf, die miteinander mittels eines Kommunikationsnetzes gekoppelt

sind. In dem Server-Computer ist ein erstes statistisches Abbild, welches für eine erste Datenbank gebildet ist, gespeichert. Das erste statistische Abbild beschreibt die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente. Der Client-Computer ist derart eingerichtet, dass mit ihm eine Weiterverarbeitung, beispielsweise eine Analyse, des von dem Server-Computer über das Kommunikationsnetz zu dem Client-Computer übertragenen ersten statistischen Abbildes möglich ist.

Bei einem Verfahren zum rechnergestützten Bilden eines statistischen Modells einer Datenbank, welche eine Vielzahl von Datenelementen aufweist, kann ein so genanntes EM-Lernverfahren (Expectation Maximisation-Lernverfahren) auf die Datenelemente durchgeführt werden, sowie auch alternativ andere Lernverfahren. Die Struktur des gemeinsamen (alle Felder in der Datenbank umfassenden) Wahrscheinlichkeitsmodells kann im Rahmen des allgemeinen Formalismus der Bayesianischen Netze (synonym auch Kausale Netze oder allgemeine Graphische Probabilistische Netze) festgelegt werden. Hierbei wird die Struktur durch einen gerichteten Graphen festgelegt. Der gerichtete Graph weist Knoten und die Knoten miteinander in Bezug setzende Kanten auf, wobei die Knoten vorgebbare Dimensionen des Modells bzw. des Abbildes entsprechend den in der Datenbank vorhandenen Werten beschreiben. Einige Knoten können dabei auch nicht beobachtbaren Größen (so genannten latenten Variablen, wie sie beispielsweise in [1] beschrieben sind) entsprechen. Im Rahmen eines allgemeinen EM-Lernverfahrens werden fehlende oder nicht beobachtbare Größen durch Erwartungswerte oder erwartete Verteilungen ersetzt. Im Rahmen des erfindungsgemäßen verbesserten EM-Lernverfahrens werden nur die Erwartungswerte ermittelt zu den fehlenden Größen, deren Eltern-Knoten beobachtbare Werte aus der Datenbank sind.

Als statistisches Abbild wird vorzugsweise ein statistisches Modell verwendet.

Unter einem statistischen Modell ist in diesem Zusammenhang jedes Modell zu verstehen, das alle statistischen Zusammenhänge bzw. die gemeinsame Häufigkeitsverteilung der Daten einer Datenbank darstellt (exakt oder approximativ), beispielsweise ein Bayesianisches (oder Kausales) Netz, ein Markov Netz oder allgemein ein Graphisches Probabilistisches Modell, ein „Latent Variabel Model", ein statistisches Clustering-Modell oder ein trainiertes künstliches Neuronales Netz. Das statistische Modell kann somit als ein vollständiges, exaktes oder approximatives Abbild der Statistik der Datenbank aufgefasst werden.

Im Zusammenhang der Weiterverarbeitung des statistischen Modells durch den Client-Computer bedeutet dies, dass eine Analyse nicht wie gemäß dem Stand der Technik basierend auf den Datenelementen der Datenbank selbst oder basierend auf einem OLAP-Werkzeug erfolgt. Stattdessen werden alle gewünschten (bedingten) Wahrscheinlichkeitsverteilungen aus dem gemeinsamen Wahrscheinlichkeitsmodell, dem statistischen Modell, ermittelt.

Diese erfindungsgemäße Vorgehensweise hat insbesondere die folgenden Vorteile:

- Verglichen mit der Datenbank selbst ist das statistische Modell sehr klein, da das statistische Modell ein komprimiertes Abbild der Statistik der Datenbank ist (nicht der einzelnen Einträge in der Datenbank), vergleichbar einem gemäß dem JPEG-Standard komprimiertem digitalen Bild, welches ein komprimiertes aber approximatives Abbild des digitalen Bildes darstellt;
- Das statistische Modell selbst kann mit wesentlich geringerem Hardware-Aufwand sehr schnell evaluiert werden.

Je nach verwendetem Verfahren zum Trainieren des statistischen Modells kann eine erhebliche Kompression der

Datenbank erzielt werden. Unter Verwendung eines in der erzielbaren Kompression skalierbaren Lernverfahrens wurde eine Kompression von bis zu einem Faktor 1000 erreicht, wobei die in dem statistischen Modell enthaltene Information qualitativ ausreichend war. Die komprimierten statistischen Modelle lassen sich somit sehr einfach beispielsweise mittels elektronischer Post (E-Mail), FTP (File Transfer Protocol) oder anderer Kommunikationsprotokolle zur Datenübertragung von dem Server-Computer zu dem Client-Computer übertragen. Das übertragene statistische Modell kann somit clientseitig zur nachfolgenden statistischen Analyse genutzt werden.

Der Server-Computer und der Client-Computer können über ein beliebiges Kommunikationsnetz, beispielsweise über ein Festnetz oder über ein Mobilfunknetz miteinander zur Übertragung des statistischen Modells gekoppelt sein.

Die Erfindung ist zum Einsatz in jedem Bereich geeignet, in dem es wünschenswert ist, nicht die gesamten Daten einer großen Datenbank zu übertragen, sondern nur eine möglichst geringe Datenmenge zu übertragen bei Erhalt eines möglichst großen Informationsgehalts der übertragenen Daten hinsichtlich der Datenbank, die von den übertragenen Daten beschrieben werden.

Ein Vorteil der Erfindung ist insbesondere darin zu sehen, dass es ermöglicht wird, in einem hohen Maße die Vertraulichkeit von individuellen Einträgen in die Datenbank zu gewährleisten, da nicht alle Datenelemente der Datenbank selbst übertragen werden, sondern nur eine statistische Repräsentation der Datenelemente der Datenbank, womit clientseitig eine statistische Analyse der Datenbank möglich wird, ohne dass clientseitig die konkreten, möglicherweise geheim zu haltenden Daten verfügbar sind.

Ferner kann ein Betreiber beispielsweise einer technischen Anlage die statistischen Inhalte der von ihm geführten

Datenbank einem Nutzer eines Client-Computers unkompliziert und in der Regel ohne Verletzung von Datenschutzrichtlinien, beispielsweise mittels eines auf dem Server-Computer installierten Web-Servers bereitgestellt werden, in welchem Fall die statistischen Modelle mittels eines auf einem Client-Computer installierten Web-Browser-Programms abgerufen werden können.

Die Erfindung kann mittels Software, das heißt mittels eines Computerprogramms, in Hardware, das heißt mittels einer speziellen elektronischen Schaltung, oder in beliebig hybrider Form, das heißt teilweise in Software und teilweise in Hardware, realisiert werden.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die folgenden Ausgestaltungen der Erfindung betreffen die Verfahren und die Computer-Anordnung.

Gemäß einer Ausgestaltung der Erfindung ist es vorgesehen, unter Verwendung des ersten statistischen Modells und Datenelementen einer in dem Client-Computer gespeicherten zweiten Datenbank ein statistisches Gesamt-Modell bzw. ein statistisches Gesamt-Abbild zu bilden, welches zumindest einen Teil der in dem ersten statistischen Abbild und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

Gemäß einer anderen Ausgestaltung der Erfindung ist es vorgesehen, für eine zweite Datenbank ein zweites statistisches Abbild bzw. ein zweites statistisches Modell zu bilden, welches die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert. Das zweite statistische Abbild wird über das Kommunikationsnetz zu dem Client-Computer übertragen und unter Verwendung des ersten statistischen Abbildes und des

zweiten statistischen Abbildes wird von dem Client-Computer ein statistisches Gesamt-Abbild gebildet, welches zumindest einen Teil der in dem ersten statistischen Abbild und in dem zweiten statistischen Abbild enthaltenen statistischen Information aufweist.

Diese Ausgestaltungen der Erfindung tragen beispielsweise folgendem allgemeinen erfindungsgemäßen Szenario Rechnung, dass fast jeder Vorgang in einem Unternehmen, insbesondere auch jeder Kundenkontakt und jede Bestellung und Auslieferung eines Produktes mit Rechnerunterstützung abläuft. In diesem Zusammenhang werden üblicherweise die Vorgänge in dem Unternehmen oder jede Aktion eines Kunden im Detail in einer Protokolldatei aufgezeichnet, beispielsweise im Rahmen von so genannten Customer Relationship Management Systemen (CRM-Systemen) oder im Rahmen von Supply Chain Management Systemen. Die protokollierten Daten stellen für viele Unternehmen ein erhebliches Vermögen dar. Dementsprechend zeigt sich ein Trend der Unternehmen, dass sie ihre Daten, beispielsweise Daten über Kunden, in „Wissen über Kunden" umsetzen. Es hat sich jedoch gezeigt, dass die in einem Unternehmen vorhandenen Informationen beispielsweise über einen Kunden (aber auch über den Betrieb einer technischen Anlage oder ähnlichem) nur sehr einseitig ist. Häufig fehlen wesentliche Attribute aller oder einzelner Kunden oder technischen Anlagen, die z.B. ein Zielgruppen-gerechtes Marketing, allgemein eine qualitativ hochwertige Datenauswertung, erst ermöglichen. Ein Beispiel im Rahmen der Kundeninformation ist in dem Alter des Kunden zu sehen oder in deren Familienstand sowie die Anzahl der Kinder. Es hat sich jedoch herausgestellt, dass bei Zusammenführen der Information mehrerer Datenbanken, seien es Kundendatenbanken oder auch Datenbanken mit Informationen über technische Prozesse, ein erheblich genaueres und vollständigeres „Bild" (im Fall des Marketings, ein „Kundenbild") ergeben. Die gemeinsame Nutzung der Datenbanken bzw. des Wissens mehrerer Unternehmen würde somit für die nachfolgende Auswertung eine

erhebliche Verbesserung ermöglichen. Der Austausch von Daten über Unternehmensgrenzen hinweg stellt aber aus folgenden Gründen keine zufrieden stellende Lösung für das oben beschriebene Problem dar:

- Unternehmen sind üblicherweise nicht bereit, Details über ihre Kunden oder ihre technischen Prozesse an andere Unternehmen weiterzugeben. Der Kundenstamm eines Unternehmens und damit die Detail-Daten über die Kunden stellen häufig ein wesentliches Unternehmensvermögen dar.
- Ein Austausch der Datenbankdaten bedeutet technisch auch, dass große Mengen an Daten übertragen und gespeichert werden müssen.
- Aus datenschutzrechtlichen Gründen sind dem Austausch von Datenbankdaten, insbesondere von personenbezogenen Daten enge Grenzen gesetzt.
- Selbst wenn Daten zwischen zwei Unternehmen ausgetauscht werden, entsteht ohne zusätzliche Maßnahmen zunächst nur für die Kunden, die in beiden Unternehmen bekannt sind, ein verbessertes Bild. Für Kunden, die nur in einem Unternehmen bekannt sind, bleiben die Daten und damit das Bild über diese Kunden weiterhin unvollständig.

Zusammenfassend ergeben sich somit anschaulich folgende erfindungsgemäße Aspekte:

- Das Wissen über Kunden oder Prozesse oder Anlagen, allgemein die in einer Datenbank enthaltene Information, wird so dargestellt,
  - dass es stark komprimiert und damit technisch auf einfachere Weise zwischen den Computern austauschbar ist, und
  - dass wesentliche Zusammenhänge dargestellt werden, dass jedoch Detail-Informationen nur in einem definierbaren Maß wiederzufinden sind, so dass Unternehmen mit weniger Bedenken solche Informationen austauschen und keine Datenschutzrichtlinien verletzt werden.

- Die auf diese Weise dargestellte Information aus verschiedenen Quellen (aus verschiedenen Datenbanken) kann zu einem Gesamtbild kombiniert werden, welches von allen teilnehmenden Unternehmen genutzt werden kann.

Durch die oben beschriebenen Ausgestaltungen wird es somit nunmehr möglich, unter Wahrung des Datenschutzes unter Reduzierung der benötigten Bandbreite zur Übertragung der statistischen Information, diese den Nutzern bereitzustellen, welche clientseitig die statistischen Modell zu einem Gesamtbild, dem Gesamt-Modell, zusammenführen können.

Gemäß einer anderen Ausgestaltung der Erfindung werden die statistischen Modell in unterschiedlichen Server-Computern gespeichert und jeweils von dort über ein Kommunikationsnetz zu dem Client-Computer übertragen.

In diesem Zusammenhang ist anzumerken, dass die statistischen Modelle von den Server-Computer(n) gebildet werden können, alternativ auch von anderen, möglicherweise speziell dazu eingerichteten Computern, in welchem Fall die gebildeten statistischen Modellen noch zu den Server-Computer(n), beispielsweise über ein lokales Netz, übertragen werden.

Somit können die statistischen Modelle in einem heterogenen Netz, beispielsweise im Internet, weltweit auf sehr einfache Weise bereitgestellt werden.

Mindestens eines der statistischen Modelle kann mittels eines skalierbaren Verfahrens gebildet werden, mit dem der Kompressionsgrad des statistischen Modells verglichen mit den in der jeweiligen Datenbank enthaltenen Datenelementen einstellbar ist.

Mindestens eines der statistischen Modelle kann ferner mittels eines EM-Lernverfahrens oder Varianten davon (wie sie beispielsweise in [2] beschrieben sind) oder mittels eines

gradientenbasierten Lernverfahrens gebildet werden. Beispielsweise kann das so genannte APN-Lernverfahren (Adaptive Probabilistic Network-Lernverfahren) als gradientenbasiertes Lernverfahren eingesetzt werden. Allgemein können alle Likelihood-basierten Lernverfahren oder Bayesianische Lernverfahren genutzt werden, wie sie beispielsweise in [3] beschrieben sind. Die Struktur der gemeinsamen Wahrscheinlichkeitsmodelle kann dabei in Form eines Graphischen Probabilistischen Modells (eines Bayesianischen Netzes, eines Markov Netzes oder einer Kombination davon) spezifiziert werden. Einem Spezialfall dieses allgemeinen Formalismus entsprechen so genannte Latent Variable Models oder statistische Clustering-Modelle. Darüber hinaus kann jedes Verfahren zum Lernen nicht nur der Parameter, sondern auch der Struktur Graphischer Probabilistischer Modelle aus verfügbaren Datenelementen genutzt werden, beispielsweise jedes beliebige Strukturlernverfahren [4] und [5].

Die erste Datenbank oder/und die zweite Datenbank kann/können Datenelemente aufweisen, welche mindestens eine technische Anlage beschreiben. Die die mindestens eine technische Anlage beschreibenden Datenelemente können zumindest teilweise an der technischen Anlage gemessene Werte darstellen, welche das Betriebsverhalten der technischen Anlage beschreiben.

Gemäß einer Ausgestaltung der erfindungsgemäßen Computer-Anordnung ist in dem Client-Computer eine zweite Datenbank mit Datenelementen gespeichert. Der Client-Computer weist eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und den Datenelementen der zweiten Datenbank, auf, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten statistischen Modell und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

Gemäß einer anderen Ausgestaltung der erfindungsgemäßen Computer-Anordnung ist ein zweiter Server-Computer vorgesehen, in dem ein zweites statistisches Modell, welches für eine zweite Datenbank gebildet ist, gespeichert ist, wobei das zweite statistische Modell die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert. Der Client-Computer ist mittels des Kommunikationsnetzes ebenfalls mit dem zweiten Server-Computer gekoppelt. Der Client-Computer weist eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und des zweiten statistischen Modells, auf, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten statistischen Modell und in dem zweiten statistischen Modell enthaltenen statistischen Information aufweist.

Ein Ausführungsbeispiel der Erfindung ist in den Figuren dargestellt und wird im Folgenden näher erläutert.

Es zeigen

Figur 1 ein Blockdiagramm einer Computer-Anordnung gemäß einem ersten Ausführungsbeispiel der Erfindung;

Figur 2 ein Blockdiagramm einer Computer-Anordnung gemäß einem zweiten Ausführungsbeispiel der Erfindung;

Figur 3 ein Blockdiagramm einer Computer-Anordnung gemäß einem dritten Ausführungsbeispiel der Erfindung;

Figur 4 ein Blockdiagramm einer Computer-Anordnung gemäß einem vierten Ausführungsbeispiel der Erfindung; und

Figur 5 ein Blockdiagramm einer Computer-Anordnung gemäß einem fünften Ausführungsbeispiel der Erfindung.

Fig.1 zeigt eine Computer-Anordnung 100 gemäß einem ersten Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 100 wird in einem Call Center eingesetzt. Die Computer-Anordnung 100 weist eine Vielzahl von Telefon-Endgeräten 101 auf, welche mittels Telefonleitungen 102 mit einem Call-Center-Computer 103, 104, 105 verbunden sind. In dem Call Center werden die Telefonanrufe von Mitarbeitern des Call Centers entgegengenommen und die Bearbeitung der eingehenden Telefonanrufe, insbesondere der Zeitpunkt des eingehenden Anrufs, die Dauer, eine Angabe über den Mitarbeiter, der den Anruf entgegengenommen hat, ein Angabe über den Grund des Anrufs sowie die Art der Bearbeitung des Anrufes oder auch beliebige andere Angaben werden von den Call-Center-Computern 103, 104, 105 aufgezeichnet.

Jeder Call-Center-Computer 103, 104, 105 weist auf

- eine erste Eingangs-/Ausgangsschnittstelle 106, 107, 108 zum öffentlichen Telefonnetz zur Entgegennahme des jeweiligen Telefonanrufes,
- einen Prozessor 109, 110, 111,
- einen Speicher 112, 113, 114, und
- eine zweite Eingangs-/Ausgangsschnittstelle 115, 116, 117 zu einem lokalen Netzwerk 121 des Call Centers.

Die oben genannten Komponenten innerhalb jedes Call-Center-Computers 103, 104, 105 sind mittels eines Computerbusses 118, 119, 120 miteinander gekoppelt.

Die Call-Center-Computer 103, 104, 105 sind mittels des lokalen Netzwerkes 121 mit einem Server-Computer 122 gekoppelt. Der Server-Computer 122 weist eine erste Eingangs-/Ausgangsschnittstelle 123 zu dem lokalen Netzwerk 121, einen Speicher 124, einen Prozessor 127 sowie eine zur Kommunikation über das Internet eingerichtete zweite Eingangs-/Ausgangsschnittstelle 128 auf, welche Komponenten

mittels eines Computerbusses 129 miteinander gekoppelt sind. Der Server-Computer 122 dient gemäß diesem Ausführungsbeispiel als Web-Server-Computer, wie im Folgenden noch näher erläutert wird.

Die von den Call-Center-Computern 103, 104, 105 aufgezeichneten Daten werden über das lokale Netzwerk 121 zu dem Server-Computer 122 übertragen und dort in einer Datenbank 126 gespeichert.

Ferner ist in dem Speicher 124 noch ein statistisches Modell 125 gespeichert, welches die statistischen Zusammenhänge der in der Datenbank 126 enthaltenen Datenelemente repräsentiert.

Das statistische Modell 125 wird unter Verwendung des an sich bekannten EM-Lernverfahrens gebildet. Andere alternative bevorzugt eingesetzte Verfahren zum Bilden des statistischen Modells 125 werden im Folgenden noch im Detail beschrieben.

Gemäß diesem Ausführungsbeispiel der Erfindung wird das statistische Modell 125 automatisch in regelmäßigen Zeitintervallen erneut, jeweils basierend auf den aktuellsten Daten der Datenbank 126, gebildet.

Das statistische Modell 125 wird von dem Server-Computer 122 automatisch zur Übertragung an einen oder an mehrere Client-Computer 132 bereitgestellt. Der Client-Computer 132 ist über eine zweite Kommunikationsverbindung 131, beispielsweise einer Kommunikationsverbindung, welche eine Kommunikation gemäß dem TCP/IP-Kommunikationsprotokoll ermöglicht, mit der zweiten Eingangs-/Ausgangsschnittstelle 128 des Server-Computers 122 gekoppelt.

Der Client-Computer 132 weist ebenfalls eine Eingangs-/Ausgangsschnittstelle 133, eingerichtet zur Kommunikation gemäß dem TCP/IP-Kommunikationsprotokoll auf sowie einen Prozessor 134 und einen Speicher 135.

Das in einer elektronischen Nachricht 130 von dem Server-Computer 122 an den Client-Computer 132 übertragene statistische Modell 125 wird in dem Speicher 135 des Client-Computers 132 gespeichert. Der Benutzer des Client-Computers 132 führt nunmehr eine beliebige, nutzerspezifische statistische Analyse auf das statistische Modell 125 und damit „indirekt" auf die Daten der Datenbank 126 aus, ohne dass die große Datenbank 126 an den Client-Computer 132 übertragen werden muss.

Ziel der clientseitigen statistischen Analyse kann eine Optimierung des Call Centers sein. Gemäß diesem Ausführungsbeispiel werden insbesondere Analysen hinsichtlich der Beantwortung der folgenden Fragen durchgeführt:

„Nach welcher Wartezeit in einer Warteschlange des Call Centers gibt ein Telefonanrufer üblicherweise auf?"

„Gibt es regionale oder tageszeitliche Abhängigkeiten zwischen den in dem Call Center eingehenden Telefonanrufen?"

„Zu welchem Zeitpunkt und in Abhängigkeit welcher anderen Merkmale treten welche Anfragen auf und wie viele Mitarbeiter sollten dementsprechend in dem Call Center bereitstehen?"

„Welche Routing-Strategien führen zu welchen Ergebnissen?"

Somit werden die Analysen zur Beantwortung der oben genannten Fragen von dem Benutzer des Client-Computers 132 durchgeführt. Anschließend werden dem Betreiber des Call Centers aus den Analyseergebnissen geeignete Maßnahmen zur optimierten Betreiben des Call Centers gegeben.

Fig.2 zeigt eine Computer-Anordnung 200 gemäß einem zweiten Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 200 wird im Bereich der Biotechnologie eingesetzt.

Die Computer-Anordnung 200 weist einen Server-Computer 201 auf, der einen Speicher 202, einen Prozessor 203 sowie eine zur Kommunikation gemäß den TCP/IP-Protokollen eingerichtete Eingangs-/Ausgangsschnittstelle 204 auf. Die Komponenten sind mittels eines Computerbusses 205 miteinander gekoppelt.

In dem Speicher 202 ist eine Datenbank 206 mit genetischen Sequenzen oder Aminosäuresequenzen zusammen mit den Sequenzen zugeordneten Zusatzinformationen gespeichert.

Für einen Forscher, gemäß diesem Ausführungsbeispiel ein Nutzer eines der Client-Computer 209, 210, 211, der die Eigenschaften einer (neuen) Sequenz untersucht, ist es häufig von erheblichem Interesse, Sequenzen mit gleichen oder ähnlichen Eigenschaften zu finden. Zum Durchsuchen der von dem oder den Server-Computern 201 öffentlich bereitgestellten Datenbanken stellt der Forscher mittels des über ein Kommunikationsnetz 208 mit dem Server-Computer 201 gekoppelten Client-Computers 209, 210, 211 entsprechende Such-Anfragen an den oder die Server-Computer 202. In dem Server-Computer 201 ist ein statistisches Modell 207 auf die gleiche Weise wie gemäß dem ersten Ausführungsbeispiel gebildet worden und dort gespeichert.

Jeder Client-Computer 209, 210, 211 weist auf

- eine zur Kommunikation gemäß den TCP/IP-Protokollen eingerichtete Eingangs-/Ausgangsschnittstelle 212, 213, 214,
- einen Prozessor 215, 216, 217,
- einen Speicher 218, 219, 220.

Nach erfolgter Anfrage eines Client-Computers 209, 210, 211 überträgt der Server-Computer 201 das statistische Modell 206

an den Client-Computer 209, 210, 211 in einer elektronischen Nachricht 221, 222, 223.

Nach Empfang des statistischen Modells 206 wird von dem Nutzer des Client-Computers 209, 210, 211 die von ihm zu untersuchende Sequenz mit dem statistischen Modell 206 verglichen. Ergebnis einer statistischen Analyse ist eine Angabe, wie viele ausreichend ähnliche Sequenzen in der Datenbank 206 existieren und durch welche Eigenschaften diese Sequenzen sich auszeichnen.

Fig.3 zeigt eine Computer-Anordnung 300 gemäß einem dritten Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 300 weist einen ersten Computer 301 und einen zweiten Computer 309 auf.

Der erste Computer 301 weist einen Speicher 302, einen Prozessor 303 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 304 auf, welche mittels eines Computerbusses 305 miteinander gekoppelt sind.

Der erste Computer 301 ist ein Computer eines Autohauses, welches in der in dem Speicher 302 gespeicherten Kunden-Datenbank Informationen zu Vorname und Nachname der Kunden, über Wohnort und genutzten Fahrzeugtyp, nicht jedoch über Alter, Familienstand und Gehaltseingang enthält.

Der zweite Computer 309 weist eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 310, einen Speicher 311 und einen Prozessor 312 auf, welche mittels eines Computerbusses 313 miteinander gekoppelt sind.

Der zweite Computer 309 ist ein Computer einer mit dem Autohaus kooperierenden Bank. In dem Speicher 311 des zweiten

Computers 309 ist eine zweite Kunden-Datenbank 314 gespeichert. In der zweiten Kunden-Datenbank 314 sind zu den Kunden der Bank Informationen zu Vorname und Nachname der Kunden, deren Wohnort, Familienstand, Alter und Gehaltseingang, enthalten, nicht jedoch zu dem von dem jeweiligen Kunden genutzten Fahrzeugtyp. Die Bank kann somit aus ihren gespeicherten Daten nicht ermitteln, welche Familien mit welchem Gehaltseingang typischerweise welche Autos nutzen.

Um diese Informationen zu erhalten, wäre die Zusammenlegung der beiden Kunden-Datenbanken erforderlich, was jedoch aus Datenschutz-rechtlichen Gründen nicht gestattet ist und von den beiden Firmen üblicherweise auch nicht erwünscht ist.

Erfindungsgemäß wird ausgenutzt, dass in beiden Datenbanken das Wissen jedenfalls approximativ vorhanden ist, um einen Zusammenhang beispielsweise zwischen Fahrzeugtyp und Gehaltseingang herzustellen.

In dem ersten Computer wird aus diesem Grund über die Datenbank ein statistisches Modell 306 gemäß dem EM-Lernverfahren gebildet. Das gegenüber der Datenbank komprimierte statistische Modell 306 wird zu dem zweiten Computer 309, welcher mit dem ersten Computer 301 bidirektional über das Internet 308 gekoppelt ist, in einer elektronischen Nachricht 307 übertragen.

Nach Empfang des statistischen Modells 306 wird dieses von dem zweiten Computer 309 mit der zweiten Kunden-Datenbank 314 zu einem statistischen Gesamt-Modell 315 zusammengeführt.

Zur Erläuterung des Zusammenführens des statistischen Modells 306 mit der zweiten Kunden-Datenbank 314 zu dem statistischen Gesamt-Modell 315 wird angenommen, dass zwei Partner A und B statistische Modelle austauschen wollen. Der Partner A verfügt über die Attribute W, X, Y, welche symbolisch für

eine Vielzahl beliebiger Attribute stehen. Der Partner B verfügt über die Attribute X, Y, Z. Der Partner B (gemäß diesem Ausführungsbeispiel das Autohaus) stellt dem Partner A (gemäß diesem Ausführungsbeispiel die Bank) ein statistisches Modell seiner Daten zur Verfügung, das im Folgenden mit $P_B(X,Y,Z)$ bezeichnet wird.

Ziel des Partners A ist es, aus seinen Daten zusammen mit den Daten seiner Datenbank ein statistisches Gesamt-Modell $P(W,X,Y,Z)$ zu erstellen.

Hierzu sind gemäß diesem Ausführungsbeispiel die folgenden zwei Verfahren vorgesehen:

- Der Partner A leitet aus dem statistischen Modell $P_B(X,Y,Z)$ ein bedingtes Modell $P_B(Z|X,Y)$ ab, um unter dessen Verwendung aus den ihm bekannten Informationen X und Y seiner Kunden die Eigenschaft Z seiner Kunden zu schätzen. Jeder Kunde bekommt als Wert der Variable Z (als Eintrag in einer zusätzlichen Spalte in der Datenbank) den Wert zugeordnet, der nach Maßgabe der Wahrscheinlichkeitsverteilung $P_B(Z|X,Y)$ am wahrscheinlichsten ist. Mit den auf diese Weise ergänzten Informationen W, X, Y und Z über jeden Kunden kann der Partner A nunmehr übliche statistische Analyseverfahren hinsichtlich aller vier Attribute anwenden oder ein gemeinsames statistisches Modell, das Gesamt-Modell $P_B(W,X,Y,Z)$, welches anschaulich ein virtuelles gemeinsames Datenbank-Abbild darstellt, erstellen.
- Statt für das Attribut Z den wahrscheinlichsten Wert zu ergänzen, kann es in einer alternativen Vorgehensweise sinnvoller sein, an Stelle der fehlenden Variable Z eine ganze Verteilung über seine Werte zu ergänzen und beim Erzeugen des statistischen Gesamt-Modells zu verwenden. Um in diesem Zusammenhang teilweise fehlende Information statistisch konsistent im Sinne der so genannten Likelihood eines Modells zu handhaben, wird das EM-

Lernverfahren eingesetzt. In jedem Lernschritt des iterativen EM-Lernverfahrens werden basierend auf den aktuellen Parametern Schätzungen (Expected Sufficient Statistics) über die fehlenden Größen erzeugt, die an die Stelle der fehlenden Größen treten. In dem EM-Lernverfahren kann das bedingte Modell $P_B(Z|X,Y)$ dazu verwendet werden, auch für die Variable Z Erwartungswerte oder Expected Sufficient Statistics-Werte zu ermitteln und so dieses Lernverfahren konsistent zu erweitern, um ein gemeinsames Modell verteilter Daten zu erzeugen.

Somit hat die Bank nunmehr die gesamte statistische Information verfügbar und kann entsprechende Analysen über die Daten durchführen.

In diesem Zusammenhang ist anzumerken, dass das oben beschriebene Szenario auch umgekehrt durchgeführt werden kann, d.h. dass die Bank ein statistisches Modell über die zweite Kunden-Datenbank erstellt und dieses an das Autohaus übermittelt, welches seinerseits ein statistisches Gesamt-Modell bildet. Für das Autohaus wäre es beispielsweise wünschenswert, das Alter seiner Kunden zu kennen, deren Familienstand und deren Gehaltseingang, oder jedenfalls eine Schätzung des Alters, des Familienstandes und des Gehaltseingangs. Basierend auf diesen Informationen können den Kunden somit passende Produkte viel gezielter angeboten werden, beispielsweise ist einer jungen Familie mit einem durchschnittlichen Gehaltseingang sicherlich ein anderes Auto anzubieten als einem Single mit einem hohen Gehalt.

Fig.4 zeigt eine Computer-Anordnung 400 gemäß einem vierten Ausführungsbeispiel der Erfindung.

Gemäß diesem Ausführungsbeispiel sind eine Vielzahl von n Computern 401, 413, 420 vorgesehen, die jeweils in

Computerbusses 424 miteinander gekoppelt sind. Über die Kunden-Datenbank in dem n-ten Computer 420 ist ebenfalls mittels des EM-Lernverfahrens ein statistisches Modell 425 gebildet und in dem Speicher 421 des n-ten Computers 420 gespeichert.

Die Computer 401, 413, 420 sind mittels einer jeweiligen Kommunikationsverbindung 408 mit einer Client-Computer 409.

Der Client-Computer 409 weist einen Speicher 411, einen Prozessor 412 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 410 auf, welche mittels eines Computerbusses 426 miteinander gekoppelt sind.

Die Computer 401, 413, 420 übermitteln die statistischen Modelle 406, 418, 525 an den Client-Computer 409 in jeweiligen elektronischen Nachrichten 407, 419, 427, welcher diese in dessen Speicher 410 speichert.

Im Folgenden wird zur einfacheren Darstellung das Ausführungsbeispiel nur unter Berücksichtigung des ersten statistischen Modells 406 und des zweiten statistischen Modells 418 näher erläutert. Es ist jedoch anzumerken, dass erfindungsgemäß eine beliebige Anzahl statistischer Modelle zu einem Gesamt-Modell zusammengeführt werden kann, beispielsweise mittels wiederholten Durchführens der im Folgenden beschriebenen Verfahrensschritte.

Im Unterschied zu dem dritten Ausführungsbeispiel ist es gemäß dem dritten Ausführungsbeispiel das Ziel, mehrere statistische Modelle miteinander zu einem Gesamt-Modell zu kombinieren.

Somit wird in Anlehnung an die im dritten Ausführungsbeispiel verwendeten Nomenklatur von dem Partner A ebenfalls ein statistisches Modell $P_A(W,X,Y)$ erstellt und dann werden die

Modelle $P_A(W,X,Y)$ und $P_B(X,Y,Z)$ zu einem statistischen Gesamt-Modell $P(W,X,Y,Z)$ kombiniert.

Das Gesamt-Modell $P(W,X,Y,Z)$ kann basierend auf den beiden Modellen $P_A(W,X,Y)$ und $P_B(X,Y,Z)$ definiert werden als:

- $P(W,X,Y,Z) = P_A(W,X,Y)P_B(Z|X,Y)$ oder als
- $P(W,X,Y,Z) = P_B(X,Y,Z)P_A(W|X,Y)$.

Auch Kombinationen aus beiden Vorgehensweisen sind erfindungsgemäß vorgesehen. Für den Partner A ist es am sinnvollsten, die erste obige Alternative zu wählen. Damit verfügt er über ein statistisches Gesamt-Modell 426, welches ihm in einer approximativen Weise ermöglicht, auch die Abhängigkeiten zwischen den Attributen W und Z zu analysieren (in diesem Ausführungsbeispiel die Abhängigkeit zwischen Fahrzeugtyp und Gehaltseingang). Basierend auf dem Gesamt-Modell 426 werden beispielsweise bedingte Wahrscheinlichkeitsverteilungen der Form $P(X|Z)$, z.B. eine Verteilung über oder eine Affinität zu Fahrzeugtypen bei einem gegebenen Gehaltseingang, ermittelt. Hierzu wird über die Variablen X und Y marginalisiert.

Zur Erläuterung wird angenommen, dass die Ergebnisse aus dem Gesamt-Modell 426 in einer Art eines zweistufigen Prozesses zustande kommen. Zunächst wird aus der Variable W auf die gemeinsamen Variablen X und Y basierend auf dem Modell $P_A(W,X,Y)$ geschlossen. Entsprechend allen danach erlaubten Kombinationen für die Variablen X und Y wird die bedingte Wahrscheinlichkeitsverteilung $P_B(Z|X,Y)$ (Prädiktion der Variable Z aus den Variablen X und Y) genutzt, um die Verteilung für die Variable Z zu bestimmen.

Im Unterschied zu dem Fall, in dem alle vier Variablen in einer Datenbank zu finden sind, erfolgt die Schlussfolgerung somit erfindungsgemäß indirekt; ähnlich wie bei einer Flüsterpost können dabei Informationen verloren gehen.

Im schlimmsten Fall, nämlich wenn kein Überlapp zwischen den beiden statistischen Abbildern vorliegt, dann ist auch keine Kombination der beiden Modelle möglich. Allerdings ist beispielsweise für den Fall, dass gemeinsame Variablen in den beiden Modellen vorhanden sind, möglich, ein Gesamt-Modell zu bilden, selbst wenn in den beiden Ausgangs-Datenbanken keine gemeinsamen Kunden, beispielsweise kein gemeinsamer Kundenschlüssel, vorhanden ist.

Das Gesamt-Modell 426 P(W,X,Y,Z) kann numerisch einfach gehandhabt werden, wenn der Überlapp zwischen diesen statistischen Modellen nicht zu groß ist, vorzugsweise kleiner als 10 gemeinsame Variablen. In dem Fall eines großen „Überlapp-Raums" können zusätzliche Approximationen verwendet werden, um die Ausführung der folgenden Summen zu beschleunigen, welche gemäß den obigen Ausführungsbeispielen über alle gemeinsamen Zustände der gemeinsamen Variablen X und Y gebildet werden müssen:

$$P(W|Z) \propto \sum_{x,y} P_A(W, X, Y) \cdot P_B(Z|X, Y)$$

bzw.

$$P(W, Z) = \sum_{x,y} P_A(W, X, Y) \cdot P_B(Z|X, Y).$$

Die Summen können insbesondere sehr geschickt approximiert werden basierend auf einem Ansatz durch Einführen einer zusätzlichen künstlichen Variable H und zusätzlichen bedingten Verteilungen (Tafeln im Falle diskreter Variable) $P(H|X, Y)$ und $P(Z|H)$ der Form:

$$P_{approx}(W, Z) \approx \sum_{x,y} P_A(W, X, Y) \sum_{h} P(H \mid X, Y) \cdot P_B(Z \mid H)$$

bzw.

$$P_{approx}(W, X, Y, Z) \approx P_A(W, X, Y) \sum_h P(H \mid X, Y) \cdot P_B(Z \mid H).$$

Die Struktur bzw. die Parametrisierung der bedingten Verteilungen $P(H|X, Y)$ und $P(Z|H)$ bzw. die Form der Abhängigkeit zwischen X,Y und H einerseits und H und Z andererseits wird so gewählt, dass die obigen Summen einfach auszuführen sind. Die Parameter der bedingten Verteilungen $P(H|X, Y)$ und $P(Z|H)$ werden so bestimmt, dass die approximative Gesamtverteilung $P_{approx}(W, X, Y, Z)$ möglicht gut der gewünschten Verteilung

$$P(W, X, Y, Z) = P_A(W, X, Y) \cdot P_B(Z|X, Y)$$

entspricht. Als Kostenfunktion kann hierbei insbesondere die Log-Likelihood bzw. die Kullback-Leibler-Distanz verwendet werden. Als Optimierungsverfahren bieten sich daher wiederum ein EM-Lernverfahren oder ein Gradienten-basiertes Lernverfahren an.

Das Auffinden optimaler Parameter kann und darf durchaus rechenaufwendig sein. Sobald die beiden Wahrscheinlichkeitsmodelle dann zu einem Gesamtmodell „fusioniert" sind kann das Gesamtmodell in einer sehr effizienten Art und Weise genutzt werden.

Es bietet sich insbesondere an, die Variable H als eine versteckte Variable einzuführen, also die Verteilung $P(W,X,Y,H)$ zu parametrisieren als

$$P(W, X, Y, H) = P(H) \cdot P(W, X, Y|H)$$

mit einer so genannten a priori Verteilung $P(H)$.

In dem Fall in dem das Modell $P(W,X,Y)$ bereits ursprünglich als ein Latent Variable Model parametrisiert wurde,

$$P_A(W, X, Y) = \sum_h P_A(X, Y, Z \mid H) \cdot P_A(H),$$

kann unmittelbar die bereits vorhandene latente Variable H genutzt werden.

Statt einer versteckten Variable H können auch mehrere Variablen eingeführt werden. Gleichzeitig kann auch für das Modell PB zur Vereinfachung der Numerik eine versteckte Variable K eingeführt werden. Eine Approximation des Gesamtmodells P(W,X,Y,Z) nimmt damit z.B. die Form an

$$P(W, X, Y, Z) \approx \sum_h P_A(X, Y, Z \mid H) \cdot P_A(H) \sum_k P(K \mid H) \cdot P_B(Z \mid K).$$

In diesem Modell können Summen über den Raum des Überlapps bestehend aus X und Y einfach durch bekannte Inferenzverfahren (beispielsweise das so genannte Junction-Tree-Verfahren) ausgeführt werden. Für die Fusion der beiden Modelle ist lediglich die bedingte Verteilung $P(K|H)$ durch bekannte Lernverfahren zu bestimmen.

Um das Ziel zu erreichen kleine, austauschbare jedoch aber sehr genaue „Abbilder einer Datenbank" zu generieren, sind insbesondere sehr skalierbare Lernverfahren, die hoch komprimierte Abbilder generieren, erwünscht. Gleichzeitig sollen sich die Abbilder effizient fusionieren, d.h. zusammenführen lassen, wozu man insbesondere auch sehr effizient mit fehlenden Informationen umgehen können sollte. Bekannte Lernverfahren sind insbesondere dann langsam, wenn in den Daten viele der Belegungen der Felder fehlen.

Fig.5 zeigt eine Computer-Anordnung 500 gemäß einem fünften Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 500 wird im Rahmen des Austauschs von Kundeninformation, gemäß diesem Ausführungsbeispiel im Rahmen des Austauschs von Adressinformation von Kunden, eingesetzt. Die Computer-Anordnung 500 weist einen Server-Computer 501 sowie einen oder mehrere mit diesem über ein Telekommunikationsnetz 502 verbundenen Client-Computer 503 auf.

Der Server-Computer 501 weist einen Speicher 504, einen Prozessor 505 sowie eine zur Kommunikation über das Internet eingerichtete Eingangs-/Ausgangsschnittstelle 506 auf, welche Komponenten mittels eines Computerbusses 507 miteinander gekoppelt sind. Der Server-Computer 501 dient gemäß diesem Ausführungsbeispiel als Web-Server-Computer, wie im Folgenden noch näher erläutert wird.

In dem Speicher 504 ist eine große Kunden-Datenbank 508 (insbesondere mit Adressinformation über die Kunden und das Kaufverhalten der Kunden beschreibende Information) gespeichert. Ferner ist in dem Speicher 504 noch ein statistisches Modell 509, welches von dem Server-Computer 501 über die Kunden-Datenbank 508 gebildet worden ist, gespeichert, welches die statistischen Zusammenhänge der in der Kunden-Datenbank 508 enthaltenen Datenelemente repräsentiert.

Das statistische Modell 509 wird unter Verwendung des an sich bekannten EM-Lernverfahrens gebildet. Andere alternative bevorzugt eingesetzte Verfahren zum Bilden des statistischen Modells 509 werden im Folgenden noch im Detail beschrieben.

Gemäß diesem Ausführungsbeispiel der Erfindung wird das statistische Modell 509 automatisch in regelmäßigen vorgegebenen Zeitintervallen erneut, jeweils basierend auf den aktuellsten Daten der Kunden-Datenbank 508, gebildet.

Das statistische Modell 509 wird von dem Server-Computer 501 automatisch zur Übertragung an den oder an mehrere Client-Computer 503 bereitgestellt.

Der Client-Computer 503 weist ebenfalls eine Eingangs-/Ausgangsschnittstelle 510, eingerichtet zur Kommunikation gemäß dem TCP/IP-Kommunikationsprotokoll auf sowie einen Prozessor 511 und einen Speicher 512. Die Komponenten des Client-Computers sind mittels eines Computerbusses 513 miteinander gekoppelt.

Das in einer elektronischen Nachricht 514 von dem Server-Computer 501 an den Client-Computer 503 übertragene statistische Modell 509 wird in dem Speicher 512 des Client-Computers 503 gespeichert.

In diesem Zusammenhang ist anzumerken, dass in dem statistischen Modell 509 die Details der Kunden-Datenbank 508, insbesondere die tatsächlichen Adressen der Kunden, nicht enthalten ist. Das statistische Modell 509 enthält allerdings statistische Information über das Verhalten, insbesondere über das Kaufverhalten der Kunden.

Der Benutzer des Client-Computers 503 wählt nunmehr eine für ihn interessante Gruppe von Kunden, d.h. einen für ihn interessanten Teil 515 des statistischen Modells 509, der ein für das Unternehmen des Benutzers des Client-Computers 503 interessierendes Kaufverhalten beschreibt, aus. Die Information 515 über den ausgewählten Teil des statistischen Modells 509 überträgt der Client-Computer 503 in einer zweiten elektronischen Nachricht 516 zu dem Server-Computer 501.

Unter Verwendung der empfangenen Information liest der Server-Computer 501 die mittels des Teils 515 des statistischen Modells 509 bezeichneten Kunden und die zugehörige Kunden-Detailinformation 517, insbesondere die

Adressen der Kunden, aus der Kunden-Datenbank 508 aus und übermittelt die ausgelesene Kunden-Detailinformation 517 in einer dritten elektronischen Nachricht 518 zu dem Client-Computer 503.

Auf diese Weise ist es möglich, beispielsweise für eine Marketing-Kampagne seitens des Benutzers des Client-Computers 503 gezielt die Adressen der gemäß der Kunden-Datenbank 508 für die Kampagne interessantesten Kunden des Unternehmens des Server-Computers 501 auszuwählen und von dem Server-Computer 501 zu erbitten. Ein erheblicher Vorteil ist ferner darin zu sehen, dass der Server-Computer 501 nur die Informationen an den Client-Computer 503 übermittelt, die auch an diesen übermittelt werden dürfen.

Diese Übermittlung erfolgt gemäß einer Ausgestaltung der Erfindung gegen Bezahlung. Anders ausgedrückt wird somit eine sehr effizientes so genanntes „On-Line Listbroking" realisiert.

Im Folgenden werden verschiedene skalierbare Verfahren zum Bilden eines statistischen Modells angegeben.

Zur besseren Veranschaulichung der bevorzugt eingesetzten Verbesserung eines EM-Lernverfahrens im Falle eines Naiven Bayesianischen Cluster Modells werden im Folgenden einige Grundlagen des EM-Lernverfahrens näher erläutert:

Mit $X = \{X_k, k = 1, \ldots, K\}$ wird einen Satz von K statistischen Variablen (die z.B. den Feldern einer Datenbank entsprechen können) bezeichnet.

Die Zustände der Variablen werden mit kleinen Buchstaben bezeichnet. Die Variable X_1 kann die Zustände $x_{1,1}, x_{1,2}, \ldots$ annehmen, d.h. $X_1 \in \{x_{1,i}, i = 1, \ldots, L_1\}$. L_1 ist die Anzahl der Zustände der Variable X_1. Ein Eintrag in einem Datensatz

(einer Datenbank) besteht nun aus Werten für alle Variablen, wobei $x^{\pi} \equiv \left(x_1^{\pi}, x_2^{\pi}, x_3^{\pi}, \ldots\right)$ den π-ten Datensatz bezeichnet. In dem π-ten Datensatz ist die Variable X_1 in dem Zustand x_1^{π}, die Variable X_2 in dem Zustand x_2^{π}, usw. Die Tafel hat M Einträge, d.h. $\left\{x^{\pi}, \pi = 1, \ldots, M\right\}$. Zusätzlich gibt es eine versteckte Variable oder eine Cluster-Variable, die im Folgenden mit Ω bezeichnet wird; deren Zustände sind $\{\omega_i, i = 1, \ldots, N\}$. Es gibt also N Cluster.

In einem statistischen Clustering-Modell beschreibt $P(\Omega)$ eine a priori Verteilung; $P(\omega_i)$ ist das a priori Gewicht des i-ten Clusters und $P(X|\omega_i)$ beschreibt die Struktur des i-ten Clusters oder die bedingte Verteilung der beobachtbaren (in der Datenbank enthaltenen) Größen $X = \{X_k, k = 1, \ldots, K\}$ in dem i-ten Cluster. Die a priori Verteilung und die bedingten Verteilungen für jedes Cluster parametrisieren zusammen ein gemeinsames Wahrscheinlichkeitsmodell auf $X \cup \Omega$ bzw. auf X.

In einem Naiven Bayesian Network wird vorausgesetzt, dass $p(\underline{X}|\omega_i)$ mit $\prod_{k=1}^{K} p(X_k|\omega_i)$ faktorisiert werden kann.

Im Allgemeinen wird darauf gezielt, die Parameter des Modells, also die a priori Verteilung $p(\Omega)$ und die bedingten Wahrscheinlichkeitstafeln $p(\underline{X}|\omega)$ derart zu bestimmen, dass das gemeinsame Modell die eingetragenen Daten möglichst gut widerspiegelt. Ein entsprechendes EM-Lernverfahren besteht aus einer Reihe von Iterationsschritten, wobei in jedem Iterationsschritt eine Verbesserung des Modells (im Sinne einer so genannten Likelihood) erzielt wird. In jedem Iterationsschritt werden neue Parameter $p^{neu}(...)$ basierend auf den aktuellen oder „alten" Parametern $p^{alt}(...)$ geschätzt.

Jeder EM-Schritt beginnt zunächst mit dem E-Schritt, in dem „Sufficient Statistics" in dafür bereitgehaltenen Tafeln

ermittelt werden. Es wird mit Wahrscheinlichkeitstafeln begonnen, deren Einträge mit Null-Werten initialisiert werden. Die Felder der Tafeln werden im Verlauf des E-Schrittes mit den so genannten Sufficient Statistics $S(\Omega)$ und $S(\underline{X}, \Omega)$ gefüllt, indem für jeden Datenpunkt die fehlenden Informationen (also insbesondere die Zuordnung jedes Datenpunktes zu den Clustern) durch Erwartungswerte ergänzt werden.

Um Erwartungswerte für die Clustervariable Ω zu berechnen ist die a posteriori Verteilung $p^{alt}\left(w_i|\underline{x}^\pi\right)$ zu ermitteln. Dieser Schritt wird auch als „Inferenzschritt" bezeichnet.

Im Falle eines Naive Bayesian Network ist die a posteriori Verteilung für Ω nach der Vorschrift

$$p^{alt}\left(w_i|\underline{x}^\pi\right) = \frac{1}{Z^\pi} p^{alt}(w_i) \prod_{k=1}^{K} p^{alt}\left(x_k^\pi|\omega_i\right)$$

für jeden Datenpunkt $\underline{x}^\pi$ aus den eingetragenen Informationen zu berechnen, wobei $\frac{1}{Z^\pi}$ eine vorgebbare Normierungskonstante ist.

Das Wesentliche dieser Berechnung besteht aus der Bildung des Produkts $p^{alt}\left(x_k^\pi|\omega_i\right)$ über alle $k = 1,\ldots,K$. Dieses Produkt muss in jedem E-Schritt für alle Cluster $i = 1,\ldots,N$ und für alle Datenpunkte $x^\pi, \pi = 1,\ldots,M$ gebildet werden.

Ähnlich aufwendig oft noch aufwendiger ist der Inferenzschritt für die Annahme anderer Abhängigkeitsstrukturen als einem Naive Bayesian Network, und beinhaltet damit den wesentlichen numerischen Aufwand des EM-Lernens.

Die Einträge in den Tafeln $S(\Omega)$ und $S(\underline{X}, \Omega)$ ändern sich nach Bildung des obigen Produktes für jeden Datenpunkt $x^{\pi}, \pi = 1, \ldots, M$, da $S(\omega_i)$ um $p^{alt}\left(\omega_i|\underline{x}^{\pi}\right)$ für alle i addiert wird, bzw. eine Summe alle $p^{alt}\left(\omega_i|\underline{x}^{\pi}\right)$ gebildet wird. Auf entsprechende Weise wird $S(\underline{x}, \omega_i)$ (bzw. $S(x_k, \omega_i)$ für alle Variabeln k im Falle eines Naive Bayesian Network) jeweils um $p^{alt}\left(\omega_i|\underline{x}^{\pi}\right)$ für alle Cluster i addiert. Dieses schließt zunächst den E (Expectation)-Schritt ab.

Anhand dieses Schrittes werden neue Parameter $p^{neu}(\Omega)$ und $p^{neu}(\underline{x}|\Omega)$ für das statistische Modell berechnet, wobei $p(\underline{x}|\omega_i)$ die Struktur des i-ten Cluster oder die bedingte Verteilung der in der Datenbank enthaltenden Größen $\underline{X}$ in diesem i-ten Cluster darstellt.

Im M (Maximisation)-Schritt werden unter Optimierung einer allgemeinen log Likelihood

$$L = \sum_{\pi=1}^{M} \log \sum_{i=1}^{N} p\left(\underline{x}^{\pi}|\omega_i\right)p(\omega_i) \qquad (1)$$

neue Parameter $p^{neu}(\Omega)$ und $p^{neu}(\underline{x}|\Omega)$, welche auf den bereits berechneten Sufficient Statistics basieren, gebildet.

Der M-Schritt bringt keinen wesentlichen numerischen Aufwand mehr mit sich.

Somit ist klar, dass der wesentliche Aufwand des Algorithmus in dem Inferenzschritt bzw. auf die Bildung des Produktes $\prod_{k=1}^{K} p^{alt}\left(x_k^{\pi}|\omega_i\right)$ und auf die Akkumulierung der Sufficient Statistics ruht.

Die Bildung von zahlreichen Null-Elementen in den Wahrscheinlichkeitstafeln $p^{alt}(\underline{x}|\omega_i)$ bzw. $p^{alt}(x_k|\omega_i)$ lässt sich jedoch durch geschickte Datenstrukturen und Speicherung von Zwischenergebnissen von einem EM-Schritt zum nächsten dazu ausnutzen, die Produkte effizient zu berechen.

Zum Beschleunigen des EM-Lernverfahrens wird die Bildung eines Gesamtproduktes in einem obigem Inferenzschritt, welcher aus Faktoren von a posteriori Verteilungen von Zugehörigkeitswahrscheinlichkeiten für alle eingegebene Datenpunkte besteht, wie gewöhnlich durchgeführt wird, sobald die erste Null in den dazu gehörenden Faktoren auftritt, wird die Bildung des Gesamtproduktes jedoch abgebrochen. Es lässt sich zeigen, dass für den Fall, dass in einem EM-Lernprozess ein Cluster für einen bestimmten Datenpunkt das Gewicht Null zugeordnet bekommt, dieser Cluster auch in allen weiteren EM-Schritten für diesen Datenpunkt das Gewicht Null zugeordnet bekommen wird.

Somit wird eine sinnvolle Beseitigung von überflüssigen numerischen Aufwand gewährleistet, indem entsprechende Ergebnisse von einem EM-Schritt zum nächsten zwischengespeichert werden und nur für die Cluster, die nicht das Gewicht Null haben, bearbeitet werden.

Es ergeben sich somit die Vorteile, dass aufgrund des Bearbeitungsabbruchs beim Auftreten eines Clusters mit Null Gewichten nicht nur innerhalb eines EM-Schrittes sondern auch für alle weiteren Schritte, besonders bei der Bildung des Produkts im Inferenzschritt, das EM-Lernverfahren insgesamt deutlich beschleunigt wird.

Im Verfahren zur Ermittlung einer in vorgegebenen Daten vorhandenen Wahrscheinlichkeitsverteilung werden Zugehörigkeitswahrscheinlichkeiten zu bestimmten Klassen nur bis zu einem Wert nahezu 0 in einem iterativen Verfahren berechnet, und die Klassen mit

Zugehörigkeitswahrscheinlichkeiten unterhalb eines auswählbaren Wertes im iterativen Verfahren nicht weiter verwendet.

In einer Weiterbildung des Verfahrens wird eine Reihenfolge der zu berechnenden Faktoren derart bestimmt, dass der Faktor, der zu einem selten auftretenden Zustand einer Variabel gehört, als erstes bearbeitet wird. Die selten auftretenden Werte können vor Beginn der Bildung des Produkts derart in einer geordneten Liste gespeichert werden, dass die Variabeln je nach Häufigkeit ihrer Erscheinung einer Null in der Liste geordnet sind.

Es ist weiterhin vorteilhaft, eine logarithmische Darstellung von Wahrscheinlichkeitstafeln zu benutzen.

Es ist weiterhin vorteilhaft, eine dünne Darstellung (sparse representation) der Wahrscheinlichkeitstafeln zu benutzen, z.B. in Form einer Liste, die nur die von Null verschiedenen Elemente enthält.

Ferner werden bei der Berechnung von Sufficient Statistics nur noch die Cluster berücksichtigt, die ein von Null verschiedenes Gewicht haben.

Die Cluster, die ein von Null verschiedenes Gewicht haben, können in eine Liste gespeichert werden, wobei die in der Liste gespeicherte Daten Pointer zu den entsprechenden Cluster sein können.

Das Verfahren kann weiterhin ein Expectation Maximisation Lernprozess sein, bei dem in dem Fall dass für ein Datenpunkt ein Cluster ein a posteriori Gewicht „Null" zugeordnet bekommt, dieser Cluster in allen weiteren Schritten des EM-Verfahrens für diesen Datenpunkt das Gewicht Null erhält und dass dieser Cluster in allen weiteren Schritten nicht mehr berücksichtigt werden muss.

Das Verfahren kann dabei nur noch über Cluster laufen, die ein von Null verschiedenes Gewicht haben.

I. Erstes Beispiel in einem Inferenzschritt

a) Bildung eines Gesamtproduktes mit Unterbrechung bei Nullwert

Für jeden Cluster ω_i in einem Inferenzschritt wird die Bildung eines Gesamtproduktes durchgeführt. Sobald die erste Null in den dazu gehörenden Faktoren, welche beispielsweise aus einem Speicher, Array oder einer Pointerliste herausgelesen werden können, auftritt, wird die Bildung des Gesamtproduktes abgebrochen.

Im Falle des Auftretens eines Nullwertes wird dann das zu dem Cluster gehörende a posteriori Gewicht auf Null gesetzt. Alternativ kann auch zuerst geprüft werden, ob zumindest einer der Faktoren in dem Produkt Null ist. Dabei werden alle Multiplikationen für die Bildung des Gesamtproduktes nur dann durchgeführt, wenn alle Faktoren von Null verschieden sind.

Wenn hingegen bei einem zu dem Gesamtprodukt gehörender Faktor kein Nullwert auftritt, so wird die Bildung des Produktes wie normal fortgeführt und der nächste Faktor aus dem Speicher, Array oder der Pointerliste herausgelesen und zur Bildung des Produktes verwendet.

b) Auswahl einer geeigneten Reihenfolge zur Beschleunigung der Datenverarbeitung

Eine geschickte Reihenfolge wird derart gewählt, dass, falls ein Faktor in dem Produkt Null ist, dieser Faktor mit hoher Wahrscheinlichkeit sehr bald als einer der ersten Faktoren in dem Produkt auftritt. Somit kann die Bildung des Gesamtproduktes sehr bald abgebrochen werden. Die Festlegung

der neuen Reihenfolge kann dabei entsprechend der Häufigkeit, mit der die Zustände der Variablen in den Daten auftreten, erfolgen. Es wird ein Faktor der zu einer sehr selten auftretenden Zustand einer Variable gehört, als erstes bearbeitet. Die Reihenfolge, in der die Faktoren bearbeitet werden, kann somit einmal vor dem Start des Lernverfahrens festgelegt werden, indem die Werte der Variablen in einer entsprechend geordneten Liste gespeichert werden.

c) Logarithmische Darstellung der Tafeln

Um den Rechenaufwand des oben genannten Verfahrens möglichst einzuschränken, wird vorzugsweise eine logarithmische Darstellung der Tafeln benutzt, um beispielsweise Underflow-Probleme zu vermeiden. Mit dieser Funktion können ursprünglich Null-Elemente zum Beispiel durch einen positiven Wert ersetzt werden. Somit ist eine aufwendige Verarbeitung bzw. Trennungen von Werten, die nahezu Null sind und sich voneinander durch einen sehr geringen Abstand unterscheiden, nicht weiter notwendig.

d) Umgehung von erhöhter Summierung bei der Berechnung von Sufficient Statistics

In dem Fall, dass die dem Lernverfahren zugegebenen stochastischen Variablen eine geringe Zugehörigkeitswahrscheinlichkeit zu einem bestimmten Cluster besitzen, werden im Laufe des Lernverfahrens viele Cluster das a posteriori Gewicht Null haben.

Um auch das Akkumulieren der Sufficient Statistics in dem darauf folgenden Schritt zu beschleunigen, werden nur noch solche Cluster in diesem Schritt berücksichtigt, die ein von Null verschiedenes Gewicht haben.

Dabei ist es vorteilhaft, die von Null verschiedenen Cluster in einer Liste, einem Array oder einer ähnlichen

Datenstruktur gespeichert werden, die es erlaubt, nur die von Null verschiedenen Elemente zu speichern.

II. Zweites Beispiel in einem EM Lernverfahren

a) Nicht-Berücksichtigung von Cluster mit Null-Zuordnungen für einen Datenpunkt

Insbesondere wird hier in einem EM-Lernverfahren von einem Schritt des Lernverfahrens zum nächsten Schritt für jeden Datenpunkt gespeichert, welche Cluster durch Auftreten von Nullen in den Tafeln noch erlaubt sind und welche nicht mehr.

Wo im ersten Beispiel Cluster, die durch Multiplikation mit Null ein a posteriori Gewicht Null erhalten, aus allen weiteren Berechnungen ausgeschlossen werden, um dadurch numerischen Aufwand zu sparen, werden in gemäß diesem Beispiel auch von einem EM-Schritt zum nächsten Zwischenergebnisse bezüglich Cluster-Zugehörigkeiten einzelner Datenpunkte (welche Cluster bereits ausgeschlossen bzw. noch zulässig sind) in zusätzlich notwendigen Datenstrukturen gespeichert.

b) Speichern einer Liste mit Referenzen auf relevante Cluster

Für jeden Datenpunkt oder für jede eingegebene stochastische Variable kann zunächst eine Liste oder eine ähnliche Datenstruktur gespeichert werden, die Referenzen auf die relevanten Cluster enthalten, die für diesen Datenpunkt ein von Null verschiedenes Gewicht bekommen haben.

Insgesamt werden in diesem Beispiel nur noch die erlaubten Cluster, allerdings für jeden Datenpunkt in einem Datensatz, gespeichert.

Die beiden obigen Beispiele können miteinander kombiniert werden, was den Abbruch bei „Null"-Gewichten im

Inferenzschritt ermöglicht, wobei in folgenden EM-Schritten nur noch die zulässigen Cluster nach dem zweiten Beispiel berücksichtigt werden.

Eine zweite Variante des EM-Lernverfahrens wird im Folgenden näher erläutert. Es ist darauf hinzuweisen, dass dieses Verfahren unabhängig von der Verwendung des auf diese Weise gebildeten statistischen Modells ist.

Bezugnehmend auf das oben beschriebene EM-Lernverfahren lässt sich zeigen, dass das Ergänzen fehlender Information nicht für alle Größen erfolgen muss. Erfindungsgemäß wurde erkannt, dass ein Teil der fehlenden Information „ignoriert" werden kann. Anders ausgedrückt bedeutet dies, dass nicht versucht wird, etwas über eine Zufallsvariable Y zu lernen aus Daten, in denen keine Information über die Zufallsvariable Y (einem Knoten Y) enthalten ist oder dass nicht versucht wird, etwas über die Zusammenhänge zwischen zwei Zufallsvariablen Y und X (zwei Knoten Y und X) aus Daten, in denen keine Information über die Zufallsvariablen Y und X enthalten ist.

Damit wird nicht nur der numerische Aufwand zur Durchführung des EM-Lernverfahrens wesentlich reduziert, sondern es wird ferner erreicht, dass das EM-Lernverfahren schneller konvergiert. Ein zusätzlicher Vorteil ist darin zu sehen, dass statistische Modelle mittels dieser Vorgehensweise leichter dynamisch aufbauen lassen, d.h. während des Lernprozesses können leichter Variablen (Knoten) in einem Netz, dem gerichteten Graphen, ergänzt werden.

Als anschauliches Beispiel für das erfindungsgemäße Verfahren wird angenommen, dass ein statistisches Modell Variablen enthält, die beschreiben, welche Bewertung ein Kinobesucher einem Film gegeben hat. Für jeden Film gibt es eine Variable, wobei jeder Variable eine Mehrzahl von Zuständen zugeordnet ist, wobei jeder Zustand jeweils einen Bewertungswert repräsentiert. Für jeden Kunden gibt es einen Datensatz, in

dem gespeichert ist, welcher Film welchen Bewertungswert erhalten hat. Wird ein neuer Film angeboten, so fehlen anfangs die Bewertungswerte für diesen Film. Mittels der neuen Variante des EM-Lernverfahrens ergibt sich nunmehr die Möglichkeit, das EM-Lernverfahren bis zu dem Erscheinen des neuen Films nur mit den bis dorthin bekannten Filmen durchzuführen, d.h. den neuen Film (d.h. allgemein den neuen Knoten in dem gerichteten Graphen) zunächst zu ignorieren. Erst mit Erscheinen des neuen Films wird das statistische Modell um eine neue Variable (einen neuen Knoten) dynamisch ergänzt und die Bewertungen des neuen Films werden berücksichtigt. Die Konvergenz des Verfahrens im Sinne der log Likelihood ist dabei noch immer gewährleistet; das Verfahren konvergiert sogar schneller.

Im Folgenden wird erläutert, unter welchen Bedingungen fehlende Informationen nicht berücksichtigt werden müssen.

Zur Erläuterung der Vorgehensweise wird folgende Notation verwendet. Mit H wird ein versteckter Knoten bezeichnet. Mit $\underline{O} = \{O^1, O^2, \ldots, O^M\}$ wird ein Satz von M beobachtbaren Knoten in dem gerichteten Graphen des statistischen Modells bezeichnet.

Es wird ohne Einschränkung der Allgemeingültigkeit im Folgenden ein Bayesianisches Wahrscheinlichkeitsmodell angenommen, welches gemäß folgender Vorschrift faktorisiert werden kann:

$$P(H, \underline{O}) = P(H) \prod_{\pi=1}^{M} P\left(O^{\pi} | H\right). \qquad (2)$$

Es ist in diesem Zusammenhang anzumerken, dass die beschriebene Vorgehensweise auf jedes statistische Modell anwendbar ist, und nicht auf ein Bayesianisches Wahrscheinlichkeitsmodell beschränkt ist, wie später noch im Detail dargelegt wird.

Mit Großbuchstaben werden im Weiteren Zufallsvariablen bezeichnet, wohingegen mit einem Kleinbuchstaben eine Instanz einer jeweiligen Zufallsvariable bezeichnet wird.

Es wird ein Datensatz mit N Datensatzelementen $\{\underline{o}_i, i = 1, \ldots, N\}$ angenommen, wobei für jedes Datensatzelement nur ein Teil der beobachtbaren Knoten tatsächlich beobachtet wird. Für das i-te Datensatzelement wird angenommen, dass die Knoten $\underline{X}_i$ beobachtet wird und dass die Beobachtungswerte der Knoten $\underline{Y}_i$ fehlen.

Es gilt also:

$$\underline{X}_i \cup \underline{Y}_i = \underline{O}_i . \qquad (3)$$

Es ist zu bemerken, dass für jedes Datensatzelement ein unterschiedlicher Satz von Knoten $\underline{X}_i$ beobachtet werden kann, d.h. dass gilt:

$$\underline{X}_i \neq \underline{X}_j \text{ für } i \neq j. \qquad (4)$$

Die Indizes für vorhandene Knoten werden mit κ bezeichnet, d.h. $\underline{X}_i = \{X_i^\kappa, \kappa = 1, \ldots, K_i\}$, die Indizes für nicht vorhandene Knoten werden mit λ bezeichnet, d.h. $\underline{Y}_i = \{Y_i^\lambda, \lambda = 1, \ldots, L_i\}$.

Im Falle eines Bayesianischen Netzes weist das übliche EM-Lernverfahren die folgenden Schritten auf, wie oben schon kurz dargestellt:

1) E-Schritt

Das Verfahren wird mit „leeren" Tabellen SS(H) und $SS(o^\pi, H), i = 1, \ldots, M$ (initialisiert mit „Nullen" gestartet, um darauf basierend die Schätzungen (Sufficient Statistics-Werte) zu akkumulieren. Für jedes Datensatzelement $\underline{o}_i$ werden

die a posteriori Verteilung $P(H|\underline{x}_i)$ für den versteckten Knoten H sowie die a posteriori Verbund-Verteilung $P(H, Y_i^{\pi}|\underline{x}_i)$ für jeden der nicht vorhandenen Knoten $\underline{Y}_i$ zusammen mit dem versteckten Knoten H berechnet.

Für jedes Datensatzelement i werden die Schätzungen für das statistische Modell akkumuliert gemäß folgenden Vorschriften:

$$SS(H) \quad += \quad \sum_i P(H|\underline{x}_i), \tag{5}$$

$$SS(X_i^{\kappa} = x_i^{\kappa}, H) \quad += \quad P(H|\underline{x}_i), \qquad \forall \text{ vorhandenen Knoten } X_i^{\kappa}, \tag{6}$$

$$SS(Y_i^{\lambda}, H) \quad += \quad P(H, Y_i^{\lambda}|\underline{x}_i) \quad \forall \text{ nicht vorhandenen Knoten } Y_i^{\lambda}. \tag{7}$$

Mit dem Symbol += wird die Aktualisierung, d.h. die Akkumulation der Tabellen für die Schätzungen gemäß den Werten der jeweiligen „rechten Seite" der Gleichung bezeichnet.

2) M-Schritt

In dem M-Schritt werden die Parameter für alle Knoten gemäß folgenden Vorschriften aktualisiert:

$$P(H) \propto SS(H), \tag{8}$$

$$P(O^{\pi}|H) \propto SS(O^{\pi}, H), \tag{9}$$

wobei mit dem Symbol $\propto$ angegeben wird, dass die Wahrscheinlichkeits-Tabellen beim Übertragen von SS auf P zu normieren sind.

Gemäß dem EM-Lernverfahren werden die Erwartungswerte für die nicht vorhandenen Knoten $\underline{Y}_i$ berechnet und entsprechend den

Sufficient Statistics-Werten für diese Knoten gemäß Vorschrift (7) aktualisiert.

Andererseits ist das Berechnen und Aktualisieren der Verbund-Verteilung $P(H, Y_i^\lambda | \underline{x}_i)$ für alle Knoten $Y_i^\lambda \in \underline{Y}_i$ sehr rechenaufwendig. Ferner ist das Aktualisieren der Verbund-Verteilung $P(H, Y_i^\lambda | \underline{x}_i)$ ein Grund für das langsame Konvergieren des EM-Lernverfahrens, wenn ein großer Teil an Information fehlt.

Angenommen, die Tabellen werden mit Zufallszahlen initialisiert, bevor das EM-Lernverfahren gestartet wird.

In diesem Fall entspricht die Verbund-Verteilung $P(H, Y_i^\lambda | \underline{x}_i)$ im Wesentlichen diesen Zufallszahlen im ersten Schritt. Dies bedeutet, dass die initialen Zufallszahlen in den Sufficient Statistics-Werten berücksichtigt werden gemäß dem Verhältnis der fehlenden Information bezogen auf die vorhandenen Information. Dies bedeutet, dass die initialen Zufallszahlen in jeder Tabelle nur gemäß dem Verhältnis der fehlenden Information bezogen auf die vorhandenen Information „gelöscht" werden.

Im Folgenden wird bewiesen, dass für den Fall eines Bayesianischen Netzes als statistisches Modell der Schritt gemäß Vorschrift (7) nicht notwendig ist und somit weggelassen bzw. übersprungen werden kann.

Die Log-Likelihood des Bayesianischen Netzes als statistisches Modell ist gegeben durch:

$$L[P] = \sum_{i=1}^{N} \log P(\underline{x}_i). \qquad (10)$$

Für frei vorgegebene Tabellen $B(H|\underline{x}_i)$, welche hinsichtlich dem Knoten H normiert sind, ergibt sich für die Log-Likelihood:

$$\begin{aligned} L[P] &= \sum_{i=1}^{N} B(h|\underline{x}_i) \log P(\underline{x}_i) \\ &= \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log \frac{P(\underline{x}_i, h)}{P(h|\underline{x}_i)} \\ &= \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(\underline{x}_i, h) - \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(h|\underline{x}_i) \end{aligned} \tag{11}$$

Die Summe $\sum_{h}$ bezeichnet die Summe über alle Zustände h des Knotens H.

Unter Verwendung der folgenden Definitionen für $R[P, B]$ und $H[P, B]$:

$$R[P, B] = \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(\underline{x}_i, h) \tag{12}$$

$$H[P, B] = \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(h|\underline{x}_i) \tag{13}$$

ergibt sich für die Log-Likelihood gemäß Vorschrift (11):

$$L[P] = R[P, B] - H[P, B]. \tag{14}$$

Allgemein gilt:

$$H[P, B] \leq H[P, P], \tag{15}$$

da $H[P, P] - H[P, B]$ die nicht-negative Kreuzentropie zwischen $P(h|\underline{x}_i)$ und $B(h|\underline{x}_i)$ darstellt.

In dem t-ten Schritt wird das aktuelle statistische Modell mit $P^{(t)}$ bezeichnet. Ausgehend von dem aktuellen statistischen Modell $P^{(t)}$ des t-ten Schrittes wird ein neues statistisches Modell $P^{(t+1)}$ konstruiert derart, dass gilt:

$$R\left[P^{(t+1)}, P^{(t)}\right] > R\left[P^{(t)}, P^{(t)}\right]. \tag{16}$$

Es gilt:

$$\begin{aligned} L\left[P^{(t+1)}\right] &= R\left[P^{(t+1)}, B\right] - H\left[P^{(t+1)}, B\right] \\ &= R\left[P^{(t+1)}, P^{(t)}\right] - H\left[P^{(t+1)}, P^{(t)}\right] \\ &> R\left[P^{(t)}, P^{(t)}\right] - H\left[P^{(t)}, P^{(t)}\right] \\ &= L\left[P^{(t)}\right] \end{aligned} \tag{17}$$

Die erste Zeile gilt allgemein für alle B (vergleiche Vorschrift (14)). Die zweite Zeile der Vorschrift (17) insbesondere für den Fall, dass gilt:

$$B = P^{(t)}. \tag{18}$$

Die dritte Zeile gilt aufgrund Vorschrift (15). Die letzte Zeile von Vorschrift (17) entspricht wiederum Vorschrift (14).

Somit ergibt sich, dass für den Fall $R\left[P^{(t+1)}, P^{(t)}\right] > R\left[P^{(t)}, P^{(t)}\right]$ sicher gilt:

$$L\left[P^{(t+1)}\right] > L\left[P^{(t)}\right]. \tag{19}$$

Es ist auf den Unterschied zu dem Standard-EM-Lernverfahren hinzuweisen [2], bei dem der R-Term definiert ist gemäß folgender Vorschrift:

$$R^{Standard}[P, B] = \sum_{i=1}^{N} \sum_{h, \underline{y}_i} B(\underline{y}_i, h|\underline{x}_i) \log P(\underline{x}_i, \underline{y}_i, h). \quad (20)$$

Es ist anzumerken, dass in dem Argument von P und B in der obigen Vorschrift (20) im Unterschied zu der Definition entsprechend den Vorschriften (12) und (13) auch die fehlenden Größen y auftreten.

Eine Sequenz von EM-Iterationen wird gebildet derart, dass gilt:

$$R^{Standard}[P^{(t+1)}, P^{(t)}] > R^{Standard}[P^{(t)}, P^{(t)}]. \quad (21)$$

Bei dem erfindungsgemäßen Lernverfahren wird für den Fall eines Bayesianischen Netzes eine Sequenz von EM-Iterationen derart gebildet, dass gilt:

$$R[P^{(t+1)}, P^{(t)}] > R[P^{(t)}, P^{(t)}]. \quad (16)$$

Nun wird gezeigt, dass die auf R, definiert gemäß Vorschrift (12), zu dem oben beschriebenen Lernverfahren führt, bei dem Vorschrift (7) übersprungen wird. Bei einem gegebenen aktuellen statistischen Modell $P^{(t)}$ zu einer Iteration t ist es das Ziel des Verfahrens, ein neues statistisches Modell $P^{(t+1)}$ in der Iteration t+1 zu berechnen, indem $R[P, P^{(t)}]$ bezüglich P optimiert wird. Unter Verwendung der Faktorisierung gemäß Vorschrift (2) ergibt sich:

$$R[P, P^{(t)}] = \sum_{i=1}^{N} \sum_{h} P^{(t)}(h|\underline{x}_i) \log P(h) + \sum_{i=1}^{N} \sum_{h} \sum_{\kappa=1}^{K_i} P^{(t)}(h|\underline{x}_i) \log P(x_i^{\kappa}|h). \quad (22)$$

Eine Optimierung von R in Bezug auf das Modell P führt zu dem erfindungsgemäßen Verfahren. Der erste Term führt zu der

Standard-Aktualisierung der $P(H)$ gemäß den Vorschriften (5) und (7).

Mit

$$SS(h) \equiv \sum_{i=1}^{N} P^{(t)}(h|\underline{x}_i) \log P(h) \qquad (23)$$

ergibt sich der erste Term von Vorschrift (22) zu

$$\sum_{h} \sum_{i=1}^{N} P^{(t)}(h|\underline{x}_i) \log P(h) = \sum_{h} SS(h) \log P(h), \qquad (24)$$

was im Wesentlichen der Kreuzentropie zwischen $SS(H)$ und $P(H)$ entspricht. Somit ist das optimale $P(H)$ durch $SS(H)$ gegeben. Dies entspricht dem M-Schritt gemäß Vorschrift (8).

Der zweite Term von Vorschrift (22) führt zu einer EM-Aktualisierung für die Tabellen der bedingten Wahrscheinlichkeiten $P(o^{\pi}|H)$, wie mittels der Vorschriften (6) und (9) beschrieben. Um dies zu veranschaulichen werden alle die Terme in R gesammelt, welche abhängig sind von $P(o^{\pi}|H)$. Diese Terme sind gegeben gemäß folgender Vorschrift:

$$\sum_{h} \sum_{\substack{i=1 \\ o^{\pi} \in \underline{X}_i}}^{N} P^{(t)}(h|\underline{x}_i) \log P(o^{\pi}|h). \qquad (25)$$

Die Summe $\sum_{\substack{i=1 \\ o^{\pi} \in \underline{X}_i}}^{N}$ bezeichnet die Summe über alle Datenelemente i in dem Datensatz, wobei o^{π} einer der beobachteten Knoten ist, d.h. bei dem gilt:

$$o^{\pi} \in \underline{X}_i \,. \qquad (26)$$

Zusammenfassend kann der obige Ausdruck (25) als die Kreuzentropie zwischen $P(o^{\pi}|H)$ und den Sufficient Statistics-Werten, welche gemäß Vorschrift (6) akkumuliert werden, interpretiert werden. Es ist somit nicht erforderlich, eine Aktualisierung gemäß Vorschrift (7) vorzusehen. Dies ist auf die Summe $\sum_{\substack{i=1 \\ o^{\pi} \in \underline{X}_i}}^{N}$ in Vorschrift (25) bzw. auf die Summe $\sum_{\kappa=1}^{K_i}$ in Vorschrift (22) zurückzuführen. Diese Summe berücksichtigt nur die beobachteten Knoten, im Gegensatz zu der Definition von $R^{Standard}$ gemäß Vorschrift (20), in der auch die nicht beobachteten Knoten $\underline{Y}_i$ berücksichtigt werden.

Im Folgenden wird in einem allgemeingültigeren Fall die Gültigkeit der Vorgehensweise, nicht beobachtete Knoten im Rahmen der Aktualisierung der Sufficient Statistics Tafeln nicht zu berücksichtigen, dargelegt, womit gezeigt wird, dass die Vorgehensweise nicht auf ein so genanntes Bayesianisches Netz beschränkt ist.

Es wird ein Satz von Variablen $\underline{Z} = \{Z^1, Z^2, \ldots, Z^M\}$ angenommen. Es wird ferner angenommen, dass das statistische Modell auf folgende Weise faktorisierbar ist:

$$P(\underline{Z}) = \prod_{\sigma=1}^{M} P\left(Z^{\sigma} \middle| \prod \left[Z^{\sigma}\right]\right), \qquad (27)$$

wobei mit $\prod \left[Z^{\sigma}\right]$ die „Eltern"-Knoten des Knoten Z^{σ} in dem Bayesianischen Netz bezeichnet werden. Ferner wird für jeden Knoten $\underline{Z}$ ein Datensatz $\{\underline{z}_i, i = 1, \ldots, N\}$ mit N Datensatzelementen angenommen. Wie schon oben angenommen, wird auch in diesem Fall in jedem der N Datensatzelemente ein nur ein Teil der Knoten $\underline{Z}$ beobachtet. Für das i-te

Datensatzelement wird angenommen, dass die Knoten $\underline{X}_i$ beobachtet werden; die Knoten $\overline{X}_i$ werden nicht beobachtet und es gilt:

$$\underline{Z} = \underline{X}_i \cup \overline{X}_i . \tag{28}$$

Für jedes der N Datensatzelemente werden die nicht beobachteten Knoten $\overline{X}_i$ in zwei Untermengen $\underline{H}_i$ und $\underline{Y}_i$ aufgeteilt derart, dass keiner der Knoten in den Mengen $\underline{X}_i$ und $\underline{H}_i$ ein abhängiger, d.h. nachfolgender Knoten („Kinder"-Knoten) eines Knotens in der Menge $\underline{Y}_i$ ist. Anschaulich bedeutet das, dass $\underline{Y}_i$ einem Zweig in einem Bayesianischen Netz entspricht, zu dem es keine Informationen in den Daten gibt.

Somit ergeben sich die Verbund-Verteilungen für die Knoten $\underline{X}_i$ und $\underline{H}_i$ gemäß folgender Vorschrift:

$$P(\underline{X}_i, \underline{H}_i) = \prod_{X \in \underline{X}_i} P\left(X \middle| \prod [X]\right) \prod_{H \in \underline{H}_i} P\left(H \middle| \prod [H]\right). \tag{29}$$

1) E-Schritt

Für jeden Knoten Z werden mit Null-Werten initialisierte Tabellen $SS\left(Z, \prod [Z]\right)$ gebildet bzw. bereitgestellt. Für jedes Datensatzelement i in dem Datensatz werden die a posteriori Verteilung $P\left(Z, \prod [Z] \middle| \underline{X}_i = \underline{x}_i\right)$ berechnet und die Sufficient Statistics-Werte gemäß folgender Vorschrift akkumuliert für jeden Knoten $Z \in \underline{X}_i$ und $Z \in \underline{H}_i$:

$$SS\left(Z, \prod | Z\right) \quad += \quad P\left(Z, \prod [Z] \middle| \underline{X}_i = \underline{x}_i\right). \tag{30}$$

Die Sufficient Statistics-Werte der Tabellen, welche den Knoten in $\overline{X}_i$ zugeordnet sind, werden nicht aktualisiert.

2) M-Schritt

Die Parameter (Tabellen) aller Knoten werden gemäß folgender Vorschrift aktualisiert:

$$P\left(z^{\sigma} \middle| \prod\left[z^{\sigma}\right]\right) \propto SS\left(z^{\sigma}, \prod\left[z^{\sigma}\right]\right). \qquad (31)$$

Anschaulich kann die Erfindung darin gesehen werden, dass ein breiter und einfacher (im Allgemeinen jedoch allerdings approximativer) Zugang zu der Statistik einer Datenbank (bevorzugt über das Internet) durch Bildung statistischer Modelle für die Inhalte der Datenbank geschaffen wird. Somit werden die statistischen Modelle zur „Remote Diagnose", zur so genannten „Remote Assistance" oder zum „Remote Research" über ein Kommunikationsnetz automatisch versendet. Anders ausgedrückt wird „Wissen" in Form eines statistischen Modells kommuniziert und versendet. Wissen ist häufig Wissen über die Zusammenhänge und wechselseitigen Abhängigkeiten in einer Domäne, beispielsweise über die Abhängigkeiten in einem Prozess. Ein statistisches Modell einer Domäne, welches aus den Daten der Datenbank gebildet wird, ist ein Abbild all dieser Zusammenhänge. Technisch stellen die Modelle eine gemeinsame Wahrscheinlichkeitsverteilung der Dimensionen der Datenbank dar, sind also nicht auf eine spezielle Aufgabenstellung eingeschränkt, sondern stellen beliebige Abhängigkeiten zwischen den Dimensionen dar. Komprimiert zu dem statistischen Modell lässt sich das Wissen über eine Domäne sehr einfach handhaben, versenden, beliebigen Nutzern bereitstellen, etc.

Die Auflösung des Abbildes bzw. des statistischen Modells kann entsprechend den Anforderungen des Datenschutzes oder den Bedürfnissen der Partner gewählt werden.

In diesem Dokumenten sind folgende Veröffentlichungen zitiert:

[1] Christopher M. Bishop, Latent Variable Models, M.I. Jordan (Editor), Learning in Graphical Models, Kulwer, 1998, Seiten 371 - 405

[2] M.A. Tanner, Tools for Statistical Inference, Springer, New York, 3. Auflage, 1996, Seiten 64 - 135

[3] Radford M. Neal und Geoffrey E. Hinton, A View of the EM Algorithm that Justifies Incremental, Sparse and Other Variants, M.I. Jordan (Editor), Learning in Graphical Models, Kulwer, 1998, Seiten 355 - 371

[4] D. Heckermann, Bayesian Networks for Data Mining, Data Mining and Knowledge Discovery, Seiten 79 - 119, 1997

[5] Reimar Hofmann, Lernen der Struktur nichtlinearer Abhängigkeiten mit graphischen Modellen, Dissertation an der Technischen Universität München, Verlag: dissertation.de, ISBN:3-89825-131-4

Patentansprüche

1. Verfahren zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank,

- bei dem für die erste Datenbank ein erstes statistisches Modell gebildet wird, welches die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente repräsentiert,
- bei dem das erste statistische Modell in einem Server-Computer gespeichert wird,
- bei dem das erste statistische Modell von dem Server-Computer über ein Kommunikationsnetz zu einem Client-Computer übertragen wird,
- bei dem das empfangene erste statistische Modell von dem Client-Computer weiterverarbeitet wird.

2. Verfahren gemäß Anspruch 1,
bei dem unter Verwendung des ersten statistischen Modells und Datenelementen einer in dem Client-Computer gespeicherten zweiten Datenbank ein statistisches Gesamt-Modell gebildet wird, welches zumindest einen Teil der in dem ersten statistischen Modell und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

3. Verfahren gemäß Anspruch 1,

- bei dem für eine zweite Datenbank ein zweites statistisches Modell gebildet wird, welches die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert,
- bei dem das zweite statistische Modell über das Kommunikationsnetz zu dem Client-Computer übertragen wird,
- bei dem unter Verwendung des ersten statistischen Modells und des zweiten statistischen Modells von dem Client-Computer ein statistisches Gesamt-Modell gebildet wird, welches zumindest einen Teil der in dem ersten

statistischen Modell und in dem zweiten statistischen Modell enthaltenen statistischen Information aufweist.

4. Verfahren gemäß Anspruch 3,

- bei dem das zweite statistische Modell in einem zweiten Server-Computer gespeichert wird,
- bei dem das zweite statistische Modell von dem zweiten Server-Computer über ein Kommunikationsnetz zu dem Client-Computer übertragen wird.

5. Verfahren gemäß einem der Ansprüche 1 bis 4, bei dem mindestens eines der statistischen Modelle mittels eines skalierbaren Verfahrens gebildet wird, mit dem der Kompressionsgrad des statistischen Modells verglichen mit den in der jeweiligen Datenbank enthaltenen Datenelementen einstellbar ist.

6. Verfahren gemäß einem der Ansprüche 1 bis 5, bei dem mindestens eines der statistischen Modelle mittels eines EM-Lernverfahrens oder mittels eines gradientenbasierten Lernverfahrens gebildet wird.

7. Verfahren gemäß einem der Ansprüche 1 bis 6, bei dem die erste Datenbank oder/und die zweite Datenbank Datenelemente aufweist/aufweisen, welche mindestens eine technische Anlage beschreiben.

8. Verfahren gemäß Anspruch 7, bei dem die die mindestens eine technische Anlage beschreibenden Datenelemente zumindest teilweise an der technischen Anlage gemessene Werte darstellen, welche das Betriebsverhalten der technischen Anlage beschreiben.

9. Verfahren zum rechnergestützten Bilden eines statistischen Modells einer Datenbank, welche eine Vielzahl von Datenelementen aufweist,

- bei dem ein EM-Lernverfahren auf die Datenelemente durchgeführt wird, so dass zu einem vorgebbaren gerichteten Graph statistische Zusammenhänge zwischen den Datenelementen ermittelt werden,
- wobei der gerichtete Graph Knoten und Kanten aufweist,
- wobei die Knoten vorgebbare beobachtbare Datenbank-Zustände und nicht beobachtbare Datenbank-Zustände beschreiben,
- bei dem im Rahmen des EM-Lernverfahrens nur die Erwartungswerte ermittelt werden zu den beobachtbaren Datenbank-Zuständen sowie zu den nicht beobachtbaren Datenbank-Zuständen, deren Eltern-Datenbank-Zustände beobachtbare Datenbank-Zustände sind.

10. Computer-Anordnung zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank,

- mit einem Server-Computer, in dem ein erstes statistisches Modell, welches für eine erste Datenbank gebildet ist, gespeichert ist, wobei das erste statistische Modell die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente repräsentiert,
- mit einem mit dem Server-Computer mittels eines Kommunikationsnetz gekoppelten Client-Computer, der eingerichtet ist zur Weiterverarbeitung des von dem Server-Computer über das Kommunikationsnetz zu dem Client-Computer übertragenen ersten statistischen Modells.

11. Computer-Anordnung gemäß Anspruch 10,

- bei der in dem Client-Computer eine zweite Datenbank mit Datenelementen gespeichert ist,
- wobei der Client-Computer eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und den Datenelementen der zweiten Datenbank, aufweist, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten

statistischen Modell und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

12. Computer-Anordnung gemäß Anspruch 10,

- mit einem zweiten Server-Computer, in dem ein zweites statistisches Modell, welches für eine zweite Datenbank gebildet ist, gespeichert ist, wobei das zweite statistische Modell die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert,
- wobei der Client-Computer mittels des Kommunikationsnetzes mit dem zweiten Server-Computer gekoppelt ist,
- wobei der Client-Computer eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und des zweiten statistischen Modells, aufweist, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten statistischen Modell und in dem zweiten statistischen Modell enthaltenen statistischen Information aufweist.

FIG 1

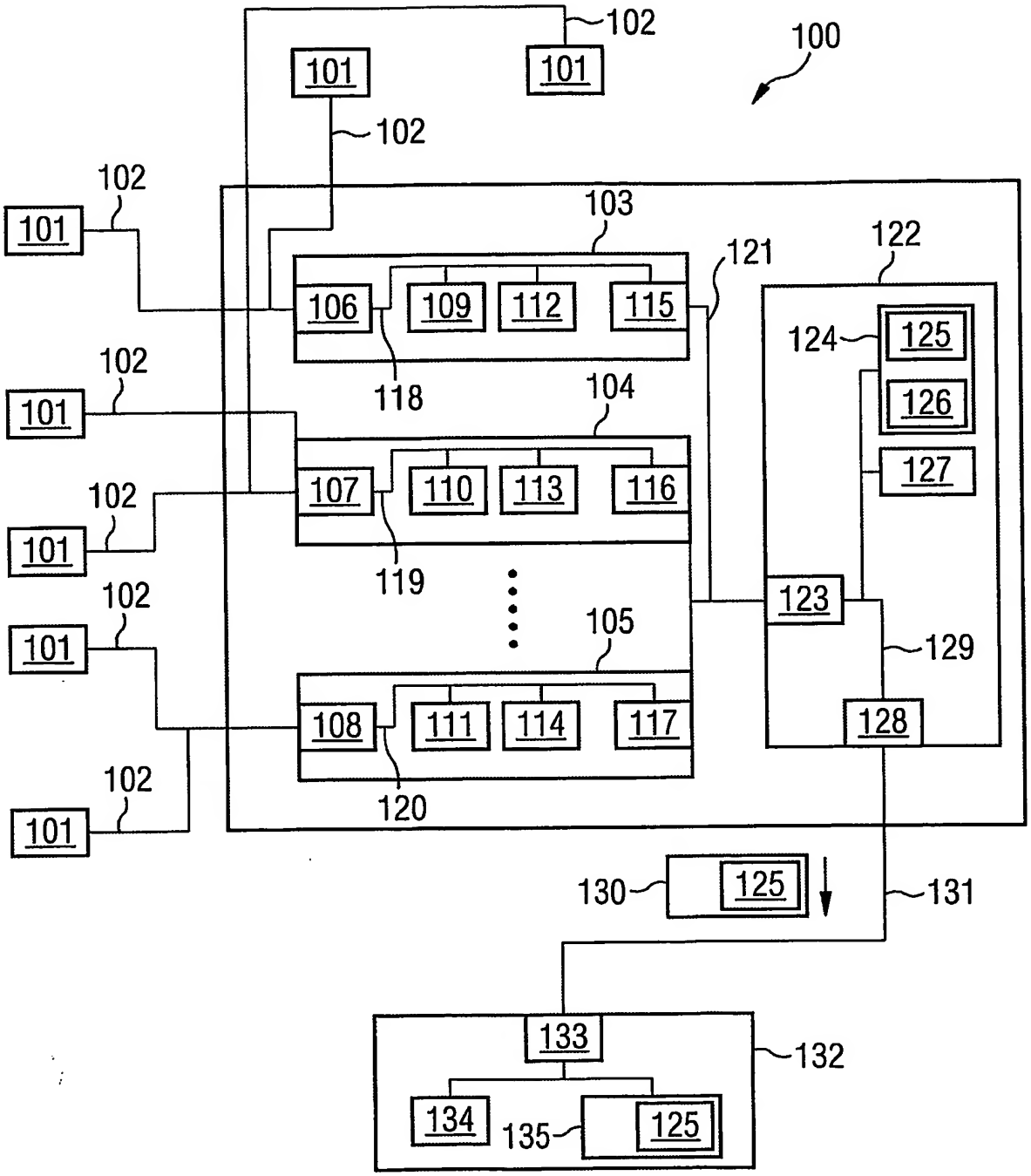

100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135

FIG 2

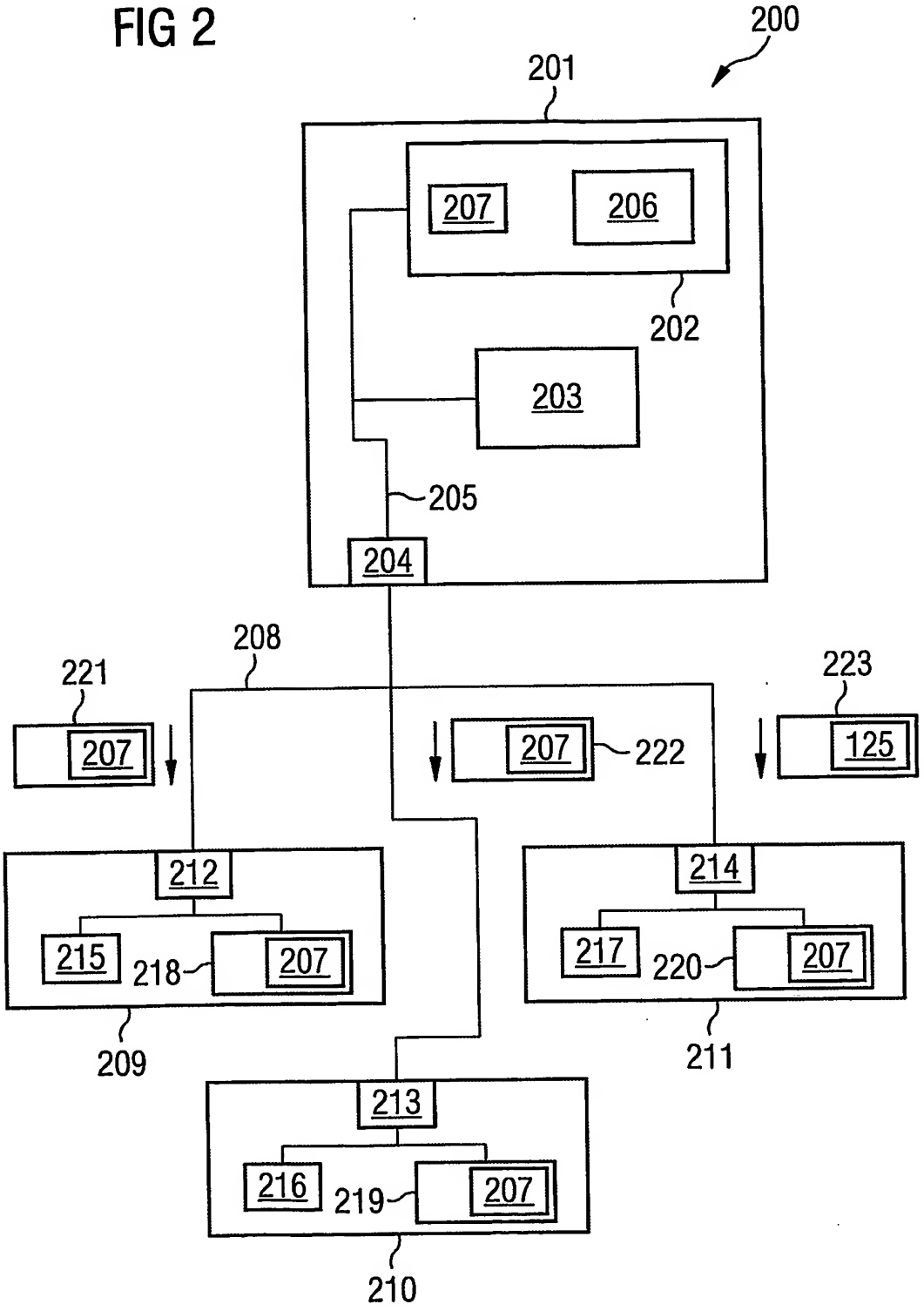

200
201
207
206
202
203
205
204
208
221
207
207
222
223
125
212
215
218
207
209
214
217
220
207
211
213
216
219
207
210

FIG 3

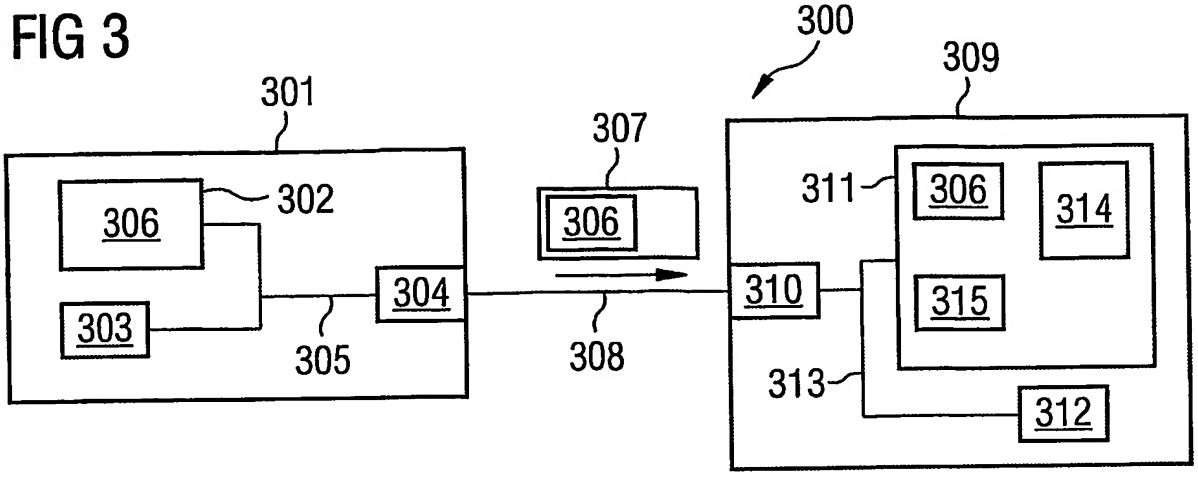

300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315

FIG 4

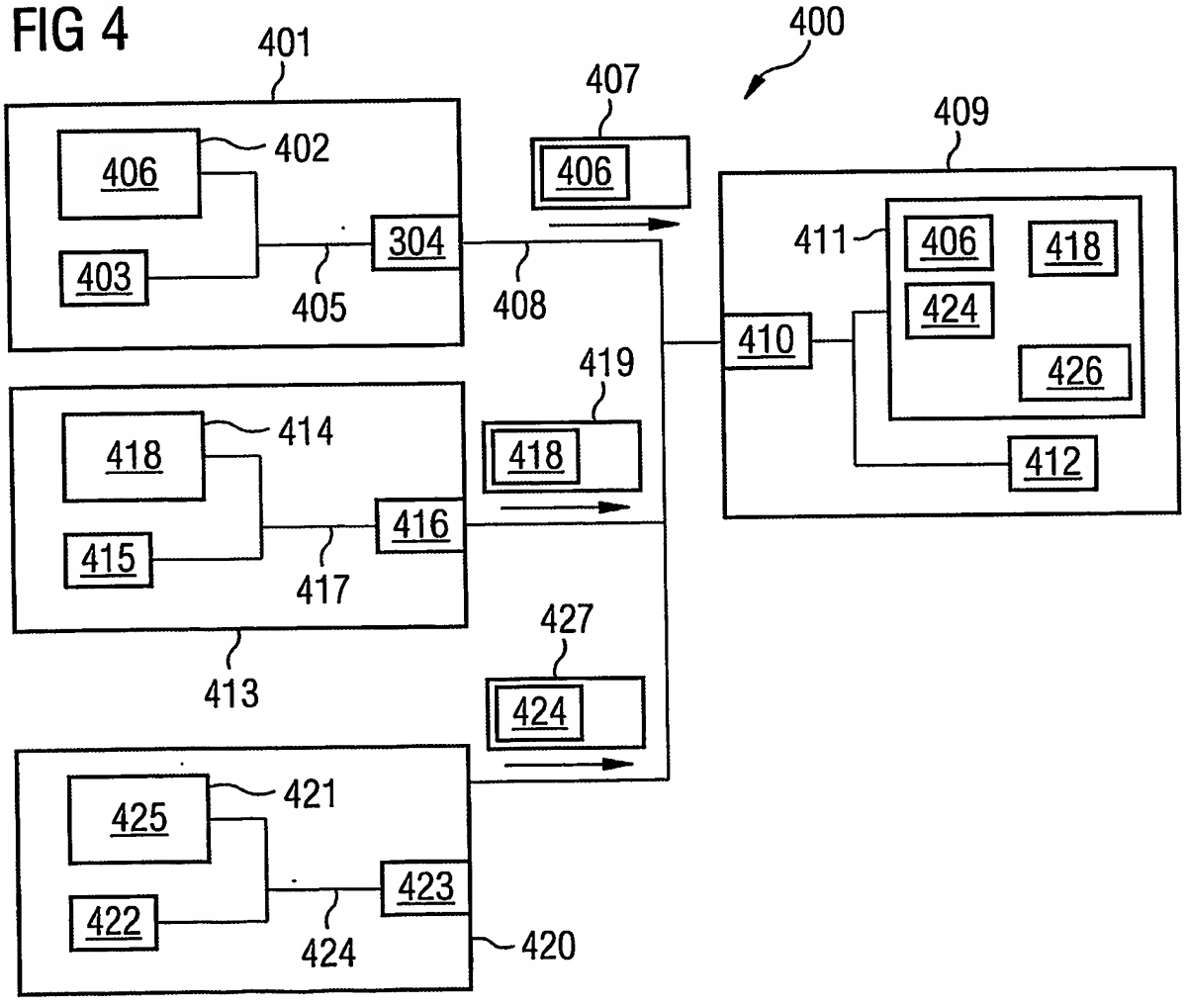
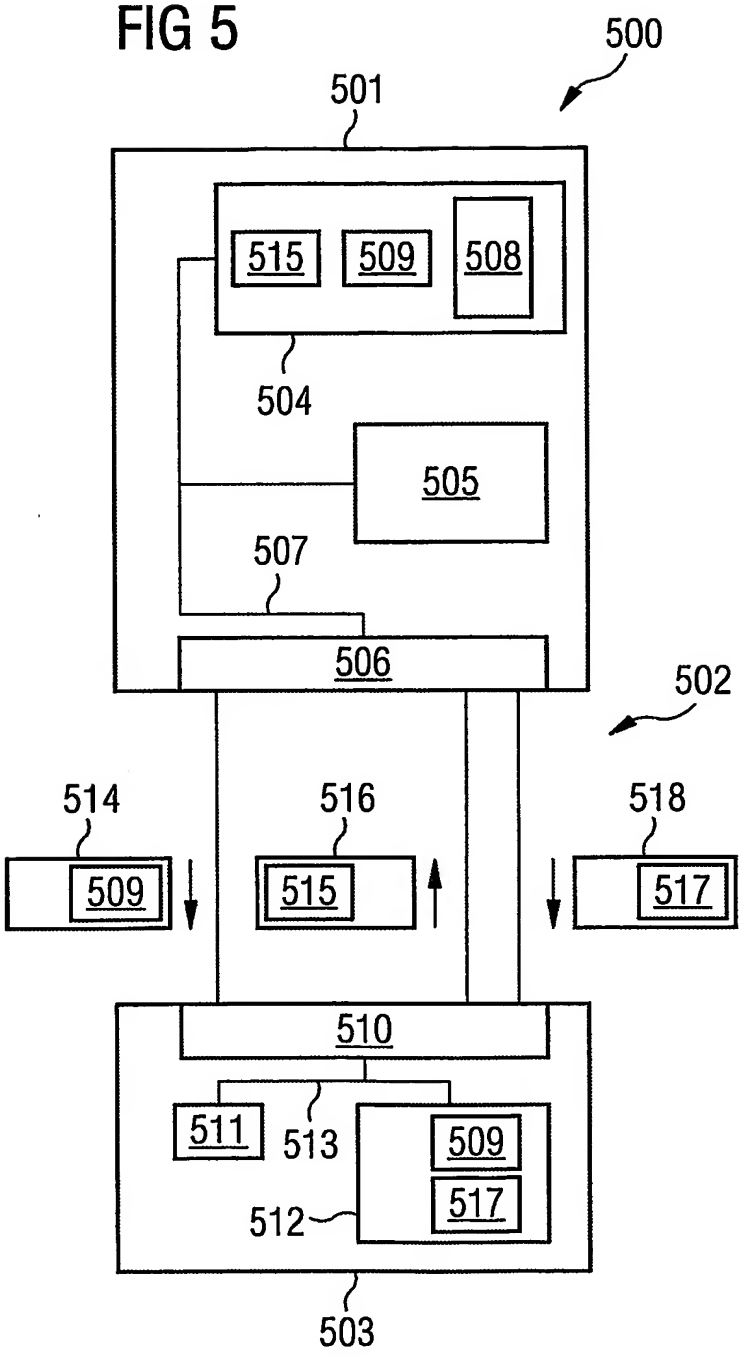

400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427

FIG 5


500
501
515
509
508
504
505
507
506
502
514
516
518
509
515
517
510
511
513
509
517
512
503

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

BERICHTIGTE FASSUNG

10/534510

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
27. Mai 2004 (27.05.2004)

PCT

(10) Internationale Veröffentlichungsnummer
WO 2004/044772 A2

(51) Internationale Patentklassifikation: G06F 17/30

(21) Internationales Aktenzeichen: PCT/EP2003/011655

(22) Internationales Anmeldedatum:
21. Oktober 2003 (21.10.2003)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
102 52 445.9 12. November 2002 (12.11.2002) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: SIEMENS AKTIENGESELLSCHAFT [DE/DE]; Wittelsbacherplatz 2, 80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: HAFT, Michael [DE/DE]; Grünlandstrasse 27, 85604 Zorneding (DE). HOFMANN, Reimar [DE/DE]; Hiltenspergerstr. 79, 80796 München (DE).

(74) Anwalt: DOKTER, Eric-Michael; Viering, Jentschura & Partner, Steinsdorfstrasse 6, 80538 München (DE).

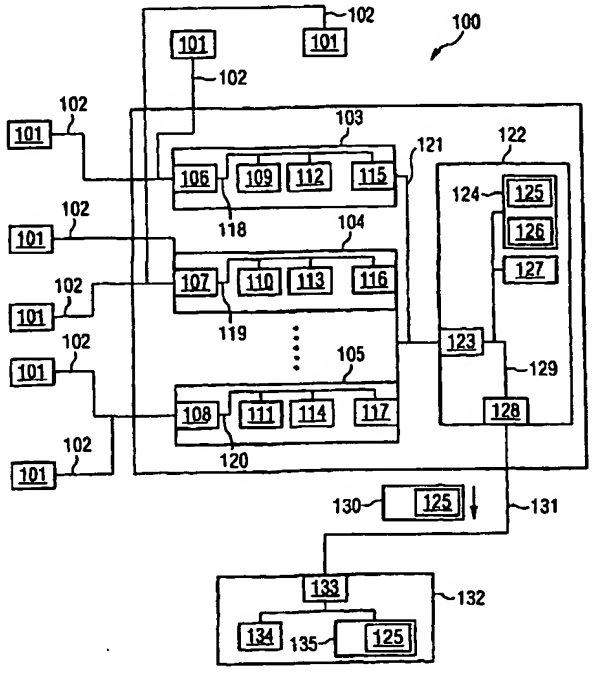
(81) Bestimmungsstaaten *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Bestimmungsstaaten *(regional)*: ARIPO Patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD AND COMPUTER CONFIGURATION FOR PROVIDING DATABASE INFORMATION OF A FIRST DATABASE AND METHOD FOR CARRYING OUT THE COMPUTER-AIDED FORMATION OF A STATISTICAL IMAGE OF A DATABASE

(54) Bezeichnung: VERFAHREN UND COMPUTER-ANORDNUNG ZUM BEREITSTELLEN VON DATENBANKINFORMATION EINER ERSTEN DATENBANK UND VERFAHREN ZUM RECHNERGESTÜTZTEN BILDEN EINES STATISTISCHEN ABBILDES EINER DATENBANK



(57) Abstract: A first statistical image is formed for the first database whereby representing statistical correlations of the data elements contained in the first database. Afterwards, the first statistical image is stored in a server computer and transmitted from this server computer over a communications network to a client computer. The received first statistical image is processed by the client computer.

(57) Zusammenfassung: Für die erste Datenbank wird ein erstes statistisches Abbild gebildet, welches die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente repräsentiert. Anschließend wird das erste statistische Abbild in einem Server-Computer gespeichert und von diesem über ein Kommunikationsnetz zu einem Client-Computer übertragen. Das empfangene erste statistische Abbild wird von dem Client-Computer weiterverarbeitet.

WO 2004/044772 A2

eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI Patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— *ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts*

(48) Datum der Veröffentlichung dieser berichtigten Fassung: 19. August 2004

(15) Informationen zur Berichtigung:
siehe PCT Gazette Nr. 34/2004 vom 19. August 2004, Section II

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

Beschreibung

Verfahren und Computer-Anordnung zum Bereitstellen von Datenbankinformation einer ersten Datenbank und Verfahren zum rechnergestützten Bilden eines statistischen Abbildes einer Datenbank

Die Erfindung betrifft ein Verfahren und eine Computer-Anordnung zum Bereitstellen von Datenbankinformation einer ersten Datenbank und ein Verfahren zum rechnergestützten Bilden eines statistischen Abbildes einer Datenbank.

Heutzutage sind kaum noch Vorgänge zu beobachten, die ohne Unterstützung eines Computers ablaufen. Häufig wird bei Einsatz eines Computers im Rahmen eines Prozesses der Prozess mittels des Computers überwacht oder zumindest prozessspezifische Daten von dem Computer aufgezeichnet und protokolliert, beispielsweise Daten über die einzelnen Prozessschritte des Prozesses und deren Ergebnisse oder Zwischenergebnisse.

Beispielsweise wird üblicherweise in einem Call Center im Detail festgehalten, wann welcher Anruf in dem Call Center eingegangen ist, wann der jeweilige eingegangene Anruf von einem Mitarbeiter des Call Centers bearbeitet wurde, zu welchem anderen Mitarbeiter des Call Centers möglicherweise weitergeleitet worden ist, etc.

Ferner werden üblicherweise in der Prozess-Automatisierung umfangreiche Protokoll-Dateien gebildet, in denen Daten über die einzelnen Prozesse gespeichert werden.

Ein drittes Anwendungsgebiet ist in der Telekommunikation zu sehen; so werden beispielsweise in den Switches eines Mobilfunknetzes Protokolldaten über den in den Switches auftretenden Datenverkehr ermittelt und gespeichert.

Schließlich werden auch in einem Webserver-Computer häufig Protokolldaten über den Datenverkehr, beispielsweise über die Zugriffshäufigkeit auf von dem Webserver-Computer bereitgestellter Information, gebildet.

Treten im Verlauf eines Prozesses Probleme auf, so wird üblicherweise der Betreiber der Anlage, auf welcher der Prozess ausgeführt wird, vor Ort versuchen, die Ursache für die aufgetretenen Probleme zu finden. Gelingt ihm das nicht, so wendet er sich meist an den Hersteller der Anlage. Herstellerseitig ist es zum Auffinden der Problemursache erforderlich, auf die protokollierten Prozessdaten, allgemein auf die aufgezeichneten Protokolldaten der Anlage zuzugreifen. Derzeit hat eine die Protokolldaten enthaltende Protokolldatei eine erhebliche Größe, häufig in der Größenordnung einiger Dutzend GByte. Eine solche Protokolldatei lässt sich aus diesem Grund nur schlecht zu dem Hersteller der Anlage, beispielsweise unter Verwendung von FTP (File Transfer Protocol) übertragen. Selbst wenn ausreichend schnelle Kommunikationsverbindungen zur Verfügung stehen, ist es für den Hersteller einer Anlage schwierig und teuer, für eine größere Anzahl von Kunden die Protokolldateien zu speichern und zu verarbeiten.

Auch in anderen Bereichen besteht der Bedarf, zu Analysezwecken große Datenmengen zu übertragen, beispielsweise überall dort, wo große Datenbanken öffentlich zugänglich sind, um der Öffentlichkeit das Forschen unter Verwendung der Datenbankdaten zu ermöglichen. Die Datenbankdaten können Daten sein aus (öffentlichen) Forschungsprojekten (beispielsweise Daten einer Gen-Datenbank oder einer Protein-Datenbank), Wetterdaten, demographische Daten, Daten, die zum Zwecke einer Rasterfahndung (in diesem Fall nur einem begrenzten Kreis befugter Nutzer) zur Verfügung gestellt werden sollen. Insbesondere der Bereich der Biotechnologie ist heutzutage von erheblichem Interesse.

Es existieren eine Vielzahl von Datenbanken in diesem Bereich.

Ferner ist es insbesondere aus Gründen der Datensicherheit häufig wünschenswert, nicht alle konkreten Informationen der Datenbankdaten weiterzugeben.

Eine bekannte Möglichkeit, Informationen einer Datenbank über ein Kommunikationsnetz von einem Server-Computer einem Client-Computer bereitzustellen, besteht darin, Diagnose- oder Statistik-Werkzeuge zur Analyse der in den Datenbanken enthaltenen Daten direkt serverseitig zu installieren, welche beispielsweise unter Verwendung eines Web-Servers, welcher auf dem Server-Computer installiert ist und eines auf einem Client-Computer installierten Web-Browser-Programms genutzt werden können. Hierfür können so genannte OLAP-Werkzeuge (On-Line Analytical Processing-Werkzeuge) eingesetzt werden, deren Betrieb allerdings sehr aufwendig und teuer ist. Bei einigen OLAP-Werkzeugen ist die zu verarbeitende Datenmenge sogar schon so groß geworden, so dass die OLAP-Werkzeuge versagen.

Ferner ist es für den Betreiber einer Anlage sehr unbequem und teuer, diese Werkzeuge serverseitig zu betreiben, da das unmittelbare Interesse an der Information ja bei dem Nutzer des Client-Computers liegt und häufig der Betreiber der Anlage nicht bereit ist, die zusätzlichen Kosten für die Bereitstellung und Wartung des Server-Computers und der OLAP-Werkzeuge zu tragen.

Weiterhin ist bei einer großen Anzahl von Client-Computern und einer großen Zahl von Anfragen an den Server-Computer die Beantwortung aller Anfragen sehr rechenaufwendig, weshalb die Hardware des Server-Computers häufig unakzeptabel teuer ist.

Der Erfindung liegt das Problem eines effizienten Zugriffs auf den Inhalt einer Datenbank über ein Kommunikationsnetz

unter Wahrung der Vertraulichkeit der in der Datenbank enthaltenen Daten zugrunde.

Das Problem wird durch ein Verfahren und eine Computer-Anordnung zum Bereitstellen von Datenbankinformation einer ersten Datenbank sowie durch ein Verfahren zum rechnergestützten Bilden eines statistischen Modells einer Datenbank mit den Merkmalen gemäß den unabhängigen Patentansprüchen gelöst.

Das allgemeine Szenario, welches von der Erfindung adressiert wird, ist auf folgende Weise charakterisiert: An einem ersten Ort A steht eine große Menge von in einer Datenbank gespeicherten Daten zur Verfügung. An einem zweiten Ort B will jemand diese zur Verfügung stehenden Daten nutzen. Der Nutzer an dem Ort B ist weniger an einzelnen Datensätzen interessiert, sondern in erster Linie an der die Datenbankdaten charakterisierenden Statistik.

Bei einem Verfahren zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank wird für die erste Datenbank ein erstes statistisches Abbild beispielsweise in Form eines gemeinsamen Wahrscheinlichkeitsmodells gebildet. Dieses Abbild bzw. Modell repräsentiert die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente. Das erste statistische Abbild wird in einem Server-Computer gespeichert. Ferner wird das erste statistische Abbild von dem Server-Computer über ein Kommunikationsnetz zu einem Client-Computer übertragen und das empfangene erste statistische Abbild wird von dem Client-Computer weiterverarbeitet.

Eine Computer-Anordnung zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank weist einen Server-Computer und einen Client-Computer auf, die miteinander mittels eines Kommunikationsnetzes gekoppelt

sind. In dem Server-Computer ist ein erstes statistisches Abbild, welches für eine erste Datenbank gebildet ist, gespeichert. Das erste statistische Abbild beschreibt die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente. Der Client-Computer ist derart eingerichtet, dass mit ihm eine Weiterverarbeitung, beispielsweise eine Analyse, des von dem Server-Computer über das Kommunikationsnetz zu dem Client-Computer übertragenen ersten statistischen Abbildes möglich ist.

Bei einem Verfahren zum rechnergestützten Bilden eines statistischen Modells einer Datenbank, welche eine Vielzahl von Datenelementen aufweist, kann ein so genanntes EM-Lernverfahren (Expectation Maximisation-Lernverfahren) auf die Datenelemente durchgeführt werden, sowie auch alternativ andere Lernverfahren. Die Struktur des gemeinsamen (alle Felder in der Datenbank umfassenden) Wahrscheinlichkeitsmodells kann im Rahmen des allgemeinen Formalismus der Bayesianischen Netze (synonym auch Kausale Netze oder allgemeine Graphische Probabilistische Netze) festgelegt werden. Hierbei wird die Struktur durch einen gerichteten Graphen festgelegt. Der gerichtete Graph weist Knoten und die Knoten miteinander in Bezug setzende Kanten auf, wobei die Knoten vorgebbare Dimensionen des Modells bzw. des Abbildes entsprechend den in der Datenbank vorhandenen Werten beschreiben. Einige Knoten können dabei auch nicht beobachtbaren Größen (so genannten latenten Variablen, wie sie beispielsweise in [1] beschrieben sind) entsprechen. Im Rahmen eines allgemeinen EM-Lernverfahrens werden fehlende oder nicht beobachtbare Größen durch Erwartungswerte oder erwartete Verteilungen ersetzt. Im Rahmen des erfindungsgemäßen verbesserten EM-Lernverfahrens werden nur die Erwartungswerte ermittelt zu den fehlenden Größen, deren Eltern-Knoten beobachtbare Werte aus der Datenbank sind.

Als statistisches Abbild wird vorzugsweise ein statistisches Modell verwendet.

Unter einem statistischen Modell ist in diesem Zusammenhang jedes Modell zu verstehen, das alle statistischen Zusammenhänge bzw. die gemeinsame Häufigkeitsverteilung der Daten einer Datenbank darstellt (exakt oder approximativ), beispielsweise ein Bayesianisches (oder Kausales) Netz, ein Markov Netz oder allgemein ein Graphisches Probabilistisches Modell, ein „Latent Variabel Model", ein statistisches Clustering-Modell oder ein trainiertes künstliches Neuronales Netz. Das statistische Modell kann somit als ein vollständiges, exaktes oder approximatives Abbild der Statistik der Datenbank aufgefasst werden.

Im Zusammenhang der Weiterverarbeitung des statistischen Modells durch den Client-Computer bedeutet dies, dass eine Analyse nicht wie gemäß dem Stand der Technik basierend auf den Datenelementen der Datenbank selbst oder basierend auf einem OLAP-Werkzeug erfolgt. Stattdessen werden alle gewünschten (bedingten) Wahrscheinlichkeitsverteilungen aus dem gemeinsamen Wahrscheinlichkeitsmodell, dem statistischen Modell, ermittelt.

Diese erfindungsgemäße Vorgehensweise hat insbesondere die folgenden Vorteile:

- Verglichen mit der Datenbank selbst ist das statistische Modell sehr klein, da das statistische Modell ein komprimiertes Abbild der Statistik der Datenbank ist (nicht der einzelnen Einträge in der Datenbank), vergleichbar einem gemäß dem JPEG-Standard komprimiertem digitalen Bild, welches ein komprimiertes aber approximatives Abbild des digitalen Bildes darstellt;
- Das statistische Modell selbst kann mit wesentlich geringerem Hardware-Aufwand sehr schnell evaluiert werden.

Je nach verwendetem Verfahren zum Trainieren des statistischen Modells kann eine erhebliche Kompression der

Datenbank erzielt werden. Unter Verwendung eines in der erzielbaren Kompression skalierbaren Lernverfahrens wurde eine Kompression von bis zu einem Faktor 1000 erreicht, wobei die in dem statistischen Modell enthaltene Information qualitativ ausreichend war. Die komprimierten statistischen Modelle lassen sich somit sehr einfach beispielsweise mittels elektronischer Post (E-Mail), FTP (File Transfer Protocol) oder anderer Kommunikationsprotokolle zur Datenübertragung von dem Server-Computer zu dem Client-Computer übertragen. Das übertragene statistische Modell kann somit clientseitig zur nachfolgenden statistischen Analyse genutzt werden.

Der Server-Computer und der Client-Computer können über ein beliebiges Kommunikationsnetz, beispielsweise über ein Festnetz oder über ein Mobilfunknetz miteinander zur Übertragung des statistischen Modells gekoppelt sein.

Die Erfindung ist zum Einsatz in jedem Bereich geeignet, in dem es wünschenswert ist, nicht die gesamten Daten einer großen Datenbank zu übertragen, sondern nur eine möglichst geringe Datenmenge zu übertragen bei Erhalt eines möglichst großen Informationsgehalts der übertragenen Daten hinsichtlich der Datenbank, die von den übertragenen Daten beschrieben werden.

Ein Vorteil der Erfindung ist insbesondere darin zu sehen, dass es ermöglicht wird, in einem hohen Maße die Vertraulichkeit von individuellen Einträgen in die Datenbank zu gewährleisten, da nicht alle Datenelemente der Datenbank selbst übertragen werden, sondern nur eine statistische Repräsentation der Datenelemente der Datenbank, womit clientseitig eine statistische Analyse der Datenbank möglich wird, ohne dass clientseitig die konkreten, möglicherweise geheim zu haltenden Daten verfügbar sind.

Ferner kann ein Betreiber beispielsweise einer technischen Anlage die statistischen Inhalte der von ihm geführten

Datenbank einem Nutzer eines Client-Computers unkompliziert und in der Regel ohne Verletzung von Datenschutzrichtlinien, beispielsweise mittels eines auf dem Server-Computer installierten Web-Servers bereitgestellt werden, in welchem Fall die statistischen Modelle mittels eines auf einem Client-Computer installierten Web-Browser-Programms abgerufen werden können.

Die Erfindung kann mittels Software, das heißt mittels eines Computerprogramms, in Hardware, das heißt mittels einer speziellen elektronischen Schaltung, oder in beliebig hybrider Form, das heißt teilweise in Software und teilweise in Hardware, realisiert werden.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die folgenden Ausgestaltungen der Erfindung betreffen die Verfahren und die Computer-Anordnung.

Gemäß einer Ausgestaltung der Erfindung ist es vorgesehen, unter Verwendung des ersten statistischen Modells und Datenelementen einer in dem Client-Computer gespeicherten zweiten Datenbank ein statistisches Gesamt-Modell bzw. ein statistisches Gesamt-Abbild zu bilden, welches zumindest einen Teil der in dem ersten statistischen Abbild und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

Gemäß einer anderen Ausgestaltung der Erfindung ist es vorgesehen, für eine zweite Datenbank ein zweites statistisches Abbild bzw. ein zweites statistisches Modell zu bilden, welches die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert. Das zweite statistische Abbild wird über das Kommunikationsnetz zu dem Client-Computer übertragen und unter Verwendung des ersten statistischen Abbildes und des

zweiten statistischen Abbildes wird von dem Client-Computer ein statistisches Gesamt-Abbild gebildet, welches zumindest einen Teil der in dem ersten statistischen Abbild und in dem zweiten statistischen Abbild enthaltenen statistischen Information aufweist.

Diese Ausgestaltungen der Erfindung tragen beispielsweise folgendem allgemeinen erfindungsgemäßen Szenario Rechnung, dass fast jeder Vorgang in einem Unternehmen, insbesondere auch jeder Kundenkontakt und jede Bestellung und Auslieferung eines Produktes mit Rechnerunterstützung abläuft. In diesem Zusammenhang werden üblicherweise die Vorgänge in dem Unternehmen oder jede Aktion eines Kunden im Detail in einer Protokolldatei aufgezeichnet, beispielsweise im Rahmen von so genannten Customer Relationship Management Systemen (CRM-Systemen) oder im Rahmen von Supply Chain Management Systemen. Die protokollierten Daten stellen für viele Unternehmen ein erhebliches Vermögen dar. Dementsprechend zeigt sich ein Trend der Unternehmen, dass sie ihre Daten, beispielsweise Daten über Kunden, in „Wissen über Kunden" umsetzen. Es hat sich jedoch gezeigt, dass die in einem Unternehmen vorhandenen Informationen beispielsweise über einen Kunden (aber auch über den Betrieb einer technischen Anlage oder ähnlichem) nur sehr einseitig ist. Häufig fehlen wesentliche Attribute aller oder einzelner Kunden oder technischen Anlagen, die z.B. ein Zielgruppen-gerechtes Marketing, allgemein eine qualitativ hochwertige Datenauswertung, erst ermöglichen. Ein Beispiel im Rahmen der Kundeninformation ist in dem Alter des Kunden zu sehen oder in deren Familienstand sowie die Anzahl der Kinder. Es hat sich jedoch herausgestellt, dass bei Zusammenführen der Information mehrerer Datenbanken, seien es Kundendatenbanken oder auch Datenbanken mit Informationen über technische Prozesse, ein erheblich genaueres und vollständigeres „Bild" (im Fall des Marketings, ein „Kundenbild") ergeben. Die gemeinsame Nutzung der Datenbanken bzw. des Wissens mehrerer Unternehmen würde somit für die nachfolgende Auswertung eine

erhebliche Verbesserung ermöglichen. Der Austausch von Daten über Unternehmensgrenzen hinweg stellt aber aus folgenden Gründen keine zufrieden stellende Lösung für das oben beschriebene Problem dar:

- Unternehmen sind üblicherweise nicht bereit, Details über ihre Kunden oder ihre technischen Prozesse an andere Unternehmen weiterzugeben. Der Kundenstamm eines Unternehmens und damit die Detail-Daten über die Kunden stellen häufig ein wesentliches Unternehmensvermögen dar.
- Ein Austausch der Datenbankdaten bedeutet technisch auch, dass große Mengen an Daten übertragen und gespeichert werden müssen.
- Aus datenschutzrechtlichen Gründen sind dem Austausch von Datenbankdaten, insbesondere von personenbezogenen Daten enge Grenzen gesetzt.
- Selbst wenn Daten zwischen zwei Unternehmen ausgetauscht werden, entsteht ohne zusätzliche Maßnahmen zunächst nur für die Kunden, die in beiden Unternehmen bekannt sind, ein verbessertes Bild. Für Kunden, die nur in einem Unternehmen bekannt sind, bleiben die Daten und damit das Bild über diese Kunden weiterhin unvollständig.

Zusammenfassend ergeben sich somit anschaulich folgende erfindungsgemäße Aspekte:

- Das Wissen über Kunden oder Prozesse oder Anlagen, allgemein die in einer Datenbank enthaltene Information, wird so dargestellt,
  - dass es stark komprimiert und damit technisch auf einfachere Weise zwischen den Computern austauschbar ist, und
  - dass wesentliche Zusammenhänge dargestellt werden, dass jedoch Detail-Informationen nur in einem definierbaren Maß wiederzufinden sind, so dass Unternehmen mit weniger Bedenken solche Informationen austauschen und keine Datenschutzrichtlinien verletzt werden.

- Die auf diese Weise dargestellte Information aus verschiedenen Quellen (aus verschiedenen Datenbanken) kann zu einem Gesamtbild kombiniert werden, welches von allen teilnehmenden Unternehmen genutzt werden kann.

Durch die oben beschriebenen Ausgestaltungen wird es somit nunmehr möglich, unter Wahrung des Datenschutzes unter Reduzierung der benötigten Bandbreite zur Übertragung der statistischen Information, diese den Nutzern bereitzustellen, welche clientseitig die statistischen Modell zu einem Gesamtbild, dem Gesamt-Modell, zusammenführen können.

Gemäß einer anderen Ausgestaltung der Erfindung werden die statistischen Modell in unterschiedlichen Server-Computern gespeichert und jeweils von dort über ein Kommunikationsnetz zu dem Client-Computer übertragen.

In diesem Zusammenhang ist anzumerken, dass die statistischen Modelle von den Server-Computer(n) gebildet werden können, alternativ auch von anderen, möglicherweise speziell dazu eingerichteten Computern, in welchem Fall die gebildeten statistischen Modellen noch zu den Server-Computer(n), beispielsweise über ein lokales Netz, übertragen werden.

Somit können die statistischen Modelle in einem heterogenen Netz, beispielsweise im Internet, weltweit auf sehr einfache Weise bereitgestellt werden.

Mindestens eines der statistischen Modelle kann mittels eines skalierbaren Verfahrens gebildet werden, mit dem der Kompressionsgrad des statistischen Modells verglichen mit den in der jeweiligen Datenbank enthaltenen Datenelementen einstellbar ist.

Mindestens eines der statistischen Modelle kann ferner mittels eines EM-Lernverfahrens oder Varianten davon (wie sie beispielsweise in [2] beschrieben sind) oder mittels eines

gradientenbasierten Lernverfahrens gebildet werden. Beispielsweise kann das so genannte APN-Lernverfahren (Adaptive Probabilistic Network-Lernverfahren) als gradientenbasiertes Lernverfahren eingesetzt werden. Allgemein können alle Likelihood-basierten Lernverfahren oder Bayesianische Lernverfahren genutzt werden, wie sie beispielsweise in [3] beschrieben sind. Die Struktur der gemeinsamen Wahrscheinlichkeitsmodelle kann dabei in Form eines Graphischen Probabilistischen Modells (eines Bayesianischen Netzes, eines Markov Netzes oder einer Kombination davon) spezifiziert werden. Einem Spezialfall dieses allgemeinen Formalismus entsprechen so genannte Latent Variable Models oder statistische Clustering-Modelle. Darüber hinaus kann jedes Verfahren zum Lernen nicht nur der Parameter, sondern auch der Struktur Graphischer Probabilistischer Modelle aus verfügbaren Datenelementen genutzt werden, beispielsweise jedes beliebige Strukturlernverfahren [4] und [5].

Die erste Datenbank oder/und die zweite Datenbank kann/können Datenelemente aufweisen, welche mindestens eine technische Anlage beschreiben. Die die mindestens eine technische Anlage beschreibenden Datenelemente können zumindest teilweise an der technischen Anlage gemessene Werte darstellen, welche das Betriebsverhalten der technischen Anlage beschreiben.

Gemäß einer Ausgestaltung der erfindungsgemäßen Computer-Anordnung ist in dem Client-Computer eine zweite Datenbank mit Datenelementen gespeichert. Der Client-Computer weist eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und den Datenelementen der zweiten Datenbank, auf, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten statistischen Modell und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

Gemäß einer anderen Ausgestaltung der erfindungsgemäßen Computer-Anordnung ist ein zweiter Server-Computer vorgesehen, in dem ein zweites statistisches Modell, welches für eine zweite Datenbank gebildet ist, gespeichert ist, wobei das zweite statistische Modell die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert. Der Client-Computer ist mittels des Kommunikationsnetzes ebenfalls mit dem zweiten Server-Computer gekoppelt. Der Client-Computer weist eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und des zweiten statistischen Modells, auf, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten statistischen Modell und in dem zweiten statistischen Modell enthaltenen statistischen Information aufweist.

Ein Ausführungsbeispiel der Erfindung ist in den Figuren dargestellt und wird im Folgenden näher erläutert.

Es zeigen

Figur 1 ein Blockdiagramm einer Computer-Anordnung gemäß einem ersten Ausführungsbeispiel der Erfindung;

Figur 2 ein Blockdiagramm einer Computer-Anordnung gemäß einem zweiten Ausführungsbeispiel der Erfindung;

Figur 3 ein Blockdiagramm einer Computer-Anordnung gemäß einem dritten Ausführungsbeispiel der Erfindung;

Figur 4 ein Blockdiagramm einer Computer-Anordnung gemäß einem vierten Ausführungsbeispiel der Erfindung; und

Figur 5 ein Blockdiagramm einer Computer-Anordnung gemäß einem fünften Ausführungsbeispiel der Erfindung.

Fig.1 zeigt eine Computer-Anordnung 100 gemäß einem ersten Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 100 wird in einem Call Center eingesetzt. Die Computer-Anordnung 100 weist eine Vielzahl von Telefon-Endgeräten 101 auf, welche mittels Telefonleitungen 102 mit einem Call-Center-Computer 103, 104, 105 verbunden sind. In dem Call Center werden die Telefonanrufe von Mitarbeitern des Call Centers entgegengenommen und die Bearbeitung der eingehenden Telefonanrufe, insbesondere der Zeitpunkt des eingehenden Anrufs, die Dauer, eine Angabe über den Mitarbeiter, der den Anruf entgegengenommen hat, ein Angabe über den Grund des Anrufs sowie die Art der Bearbeitung des Anrufes oder auch beliebige andere Angaben werden von den Call-Center-Computern 103, 104, 105 aufgezeichnet.

Jeder Call-Center-Computer 103, 104, 105 weist auf

- eine erste Eingangs-/Ausgangsschnittstelle 106, 107, 108 zum öffentlichen Telefonnetz zur Entgegennahme des jeweiligen Telefonanrufes,
- einen Prozessor 109, 110, 111,
- einen Speicher 112, 113, 114, und
- eine zweite Eingangs-/Ausgangsschnittstelle 115, 116, 117 zu einem lokalen Netzwerk 121 des Call Centers.

Die oben genannten Komponenten innerhalb jedes Call-Center-Computers 103, 104, 105 sind mittels eines Computerbusses 118, 119, 120 miteinander gekoppelt.

Die Call-Center-Computer 103, 104, 105 sind mittels des lokalen Netzwerkes 121 mit einem Server-Computer 122 gekoppelt. Der Server-Computer 122 weist eine erste Eingangs-/Ausgangsschnittstelle 123 zu dem lokalen Netzwerk 121, einen Speicher 124, einen Prozessor 127 sowie eine zur Kommunikation über das Internet eingerichtete zweite Eingangs-/Ausgangsschnittstelle 128 auf, welche Komponenten

mittels eines Computerbusses 129 miteinander gekoppelt sind. Der Server-Computer 122 dient gemäß diesem Ausführungsbeispiel als Web-Server-Computer, wie im Folgenden noch näher erläutert wird.

Die von den Call-Center-Computern 103, 104, 105 aufgezeichneten Daten werden über das lokale Netzwerk 121 zu dem Server-Computer 122 übertragen und dort in einer Datenbank 126 gespeichert.

Ferner ist in dem Speicher 124 noch ein statistisches Modell 125 gespeichert, welches die statistischen Zusammenhänge der in der Datenbank 126 enthaltenen Datenelemente repräsentiert.

Das statistische Modell 125 wird unter Verwendung des an sich bekannten EM-Lernverfahrens gebildet. Andere alternative bevorzugt eingesetzte Verfahren zum Bilden des statistischen Modells 125 werden im Folgenden noch im Detail beschrieben.

Gemäß diesem Ausführungsbeispiel der Erfindung wird das statistische Modell 125 automatisch in regelmäßigen Zeitintervallen erneut, jeweils basierend auf den aktuellsten Daten der Datenbank 126, gebildet.

Das statistische Modell 125 wird von dem Server-Computer 122 automatisch zur Übertragung an einen oder an mehrere Client-Computer 132 bereitgestellt. Der Client-Computer 132 ist über eine zweite Kommunikationsverbindung 131, beispielsweise einer Kommunikationsverbindung, welche eine Kommunikation gemäß dem TCP/IP-Kommunikationsprotokoll ermöglicht, mit der zweiten Eingangs-/Ausgangsschnittstelle 128 des Server-Computers 122 gekoppelt.

Der Client-Computer 132 weist ebenfalls eine Eingangs-/Ausgangsschnittstelle 133, eingerichtet zur Kommunikation gemäß dem TCP/IP-Kommunikationsprotokoll auf sowie einen Prozessor 134 und einen Speicher 135.

Das in einer elektronischen Nachricht 130 von dem Server-Computer 122 an den Client-Computer 132 übertragene statistische Modell 125 wird in dem Speicher 135 des Client-Computers 132 gespeichert. Der Benutzer des Client-Computers 132 führt nunmehr eine beliebige, nutzerspezifische statistische Analyse auf das statistische Modell 125 und damit „indirekt" auf die Daten der Datenbank 126 aus, ohne dass die große Datenbank 126 an den Client-Computer 132 übertragen werden muss.

Ziel der clientseitigen statistischen Analyse kann eine Optimierung des Call Centers sein. Gemäß diesem Ausführungsbeispiel werden insbesondere Analysen hinsichtlich der Beantwortung der folgenden Fragen durchgeführt:

„Nach welcher Wartezeit in einer Warteschlange des Call Centers gibt ein Telefonanrufer üblicherweise auf?"

„Gibt es regionale oder tageszeitliche Abhängigkeiten zwischen den in dem Call Center eingehenden Telefonanrufen?"

„Zu welchem Zeitpunkt und in Abhängigkeit welcher anderen Merkmale treten welche Anfragen auf und wie viele Mitarbeiter sollten dementsprechend in dem Call Center bereitstehen?"

„Welche Routing-Strategien führen zu welchen Ergebnissen?"

Somit werden die Analysen zur Beantwortung der oben genannten Fragen von dem Benutzer des Client-Computers 132 durchgeführt. Anschließend werden dem Betreiber des Call Centers aus den Analyseergebnissen geeignete Maßnahmen zur optimierten Betreiben des Call Centers gegeben.

Fig.2 zeigt eine Computer-Anordnung 200 gemäß einem zweiten Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 200 wird im Bereich der Biotechnologie eingesetzt.

Die Computer-Anordnung 200 weist einen Server-Computer 201 auf, der einen Speicher 202, einen Prozessor 203 sowie eine zur Kommunikation gemäß den TCP/IP-Protokollen eingerichtete Eingangs-/Ausgangsschnittstelle 204 auf. Die Komponenten sind mittels eines Computerbusses 205 miteinander gekoppelt.

In dem Speicher 202 ist eine Datenbank 206 mit genetischen Sequenzen oder Aminosäuresequenzen zusammen mit den Sequenzen zugeordneten Zusatzinformationen gespeichert.

Für einen Forscher, gemäß diesem Ausführungsbeispiel ein Nutzer eines der Client-Computer 209, 210, 211, der die Eigenschaften einer (neuen) Sequenz untersucht, ist es häufig von erheblichem Interesse, Sequenzen mit gleichen oder ähnlichen Eigenschaften zu finden. Zum Durchsuchen der von dem oder den Server-Computern 201 öffentlich bereitgestellten Datenbanken stellt der Forscher mittels des über ein Kommunikationsnetz 208 mit dem Server-Computer 201 gekoppelten Client-Computers 209, 210, 211 entsprechende Such-Anfragen an den oder die Server-Computer 202. In dem Server-Computer 201 ist ein statistisches Modell 207 auf die gleiche Weise wie gemäß dem ersten Ausführungsbeispiel gebildet worden und dort gespeichert.

Jeder Client-Computer 209, 210, 211 weist auf

- eine zur Kommunikation gemäß den TCP/IP-Protokollen eingerichtete Eingangs-/Ausgangsschnittstelle 212, 213, 214,
- einen Prozessor 215, 216, 217,
- einen Speicher 218, 219, 220.

Nach erfolgter Anfrage eines Client-Computers 209, 210, 211 überträgt der Server-Computer 201 das statistische Modell 206

an den Client-Computer 209, 210, 211 in einer elektronischen Nachricht 221, 222, 223.

Nach Empfang des statistischen Modells 206 wird von dem Nutzer des Client-Computers 209, 210, 211 die von ihm zu untersuchende Sequenz mit dem statistischen Modell 206 verglichen. Ergebnis einer statistischen Analyse ist eine Angabe, wie viele ausreichend ähnliche Sequenzen in der Datenbank 206 existieren und durch welche Eigenschaften diese Sequenzen sich auszeichnen.

Fig.3 zeigt eine Computer-Anordnung 300 gemäß einem dritten Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 300 weist einen ersten Computer 301 und einen zweiten Computer 309 auf.

Der erste Computer 301 weist einen Speicher 302, einen Prozessor 303 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 304 auf, welche mittels eines Computerbusses 305 miteinander gekoppelt sind.

Der erste Computer 301 ist ein Computer eines Autohauses, welches in der in dem Speicher 302 gespeicherten Kunden-Datenbank Informationen zu Vorname und Nachname der Kunden, über Wohnort und genutzten Fahrzeugtyp, nicht jedoch über Alter, Familienstand und Gehaltseingang enthält.

Der zweite Computer 309 weist eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 310, einen Speicher 311 und einen Prozessor 312 auf, welche mittels eines Computerbusses 313 miteinander gekoppelt sind.

Der zweite Computer 309 ist ein Computer einer mit dem Autohaus kooperierenden Bank. In dem Speicher 311 des zweiten

Computers 309 ist eine zweite Kunden-Datenbank 314 gespeichert. In der zweiten Kunden-Datenbank 314 sind zu den Kunden der Bank Informationen zu Vorname und Nachname der Kunden, deren Wohnort, Familienstand, Alter und Gehaltseingang, enthalten, nicht jedoch zu dem von dem jeweiligen Kunden genutzten Fahrzeugtyp. Die Bank kann somit aus ihren gespeicherten Daten nicht ermitteln, welche Familien mit welchem Gehaltseingang typischerweise welche Autos nutzen.

Um diese Informationen zu erhalten, wäre die Zusammenlegung der beiden Kunden-Datenbanken erforderlich, was jedoch aus Datenschutz-rechtlichen Gründen nicht gestattet ist und von den beiden Firmen üblicherweise auch nicht erwünscht ist.

Erfindungsgemäß wird ausgenutzt, dass in beiden Datenbanken das Wissen jedenfalls approximativ vorhanden ist, um einen Zusammenhang beispielsweise zwischen Fahrzeugtyp und Gehaltseingang herzustellen.

In dem ersten Computer wird aus diesem Grund über die Datenbank ein statistisches Modell 306 gemäß dem EM-Lernverfahren gebildet. Das gegenüber der Datenbank komprimierte statistische Modell 306 wird zu dem zweiten Computer 309, welcher mit dem ersten Computer 301 bidirektional über das Internet 308 gekoppelt ist, in einer elektronischen Nachricht 307 übertragen.

Nach Empfang des statistischen Modells 306 wird dieses von dem zweiten Computer 309 mit der zweiten Kunden-Datenbank 314 zu einem statistischen Gesamt-Modell 315 zusammengeführt.

Zur Erläuterung des Zusammenführens des statistischen Modells 306 mit der zweiten Kunden-Datenbank 314 zu dem statistischen Gesamt-Modell 315 wird angenommen, dass zwei Partner A und B statistische Modelle austauschen wollen. Der Partner A verfügt über die Attribute W, X, Y, welche symbolisch für

eine Vielzahl beliebiger Attribute stehen. Der Partner B verfügt über die Attribute X, Y, Z. Der Partner B (gemäß diesem Ausführungsbeispiel das Autohaus) stellt dem Partner A (gemäß diesem Ausführungsbeispiel die Bank) ein statistisches Modell seiner Daten zur Verfügung, das im Folgenden mit $P_B(X,Y,Z)$ bezeichnet wird.

Ziel des Partners A ist es, aus seinen Daten zusammen mit den Daten seiner Datenbank ein statistisches Gesamt-Modell $P(W,X,Y,Z)$ zu erstellen.

Hierzu sind gemäß diesem Ausführungsbeispiel die folgenden zwei Verfahren vorgesehen:

- Der Partner A leitet aus dem statistischen Modell $P_B(X,Y,Z)$ ein bedingtes Modell $P_B(Z|X,Y)$ ab, um unter dessen Verwendung aus den ihm bekannten Informationen X und Y seiner Kunden die Eigenschaft Z seiner Kunden zu schätzen. Jeder Kunde bekommt als Wert der Variable Z (als Eintrag in einer zusätzlichen Spalte in der Datenbank) den Wert zugeordnet, der nach Maßgabe der Wahrscheinlichkeitsverteilung $P_B(Z|X,Y)$ am wahrscheinlichsten ist. Mit den auf diese Weise ergänzten Informationen W, X, Y und Z über jeden Kunden kann der Partner A nunmehr übliche statistische Analyseverfahren hinsichtlich aller vier Attribute anwenden oder ein gemeinsames statistisches Modell, das Gesamt-Modell $P_B(W,X,Y,Z)$, welches anschaulich ein virtuelles gemeinsames Datenbank-Abbild darstellt, erstellen.
- Statt für das Attribut Z den wahrscheinlichsten Wert zu ergänzen, kann es in einer alternativen Vorgehensweise sinnvoller sein, an Stelle der fehlenden Variable Z eine ganze Verteilung über seine Werte zu ergänzen und beim Erzeugen des statistischen Gesamt-Modells zu verwenden. Um in diesem Zusammenhang teilweise fehlende Information statistisch konsistent im Sinne der so genannten Likelihood eines Modells zu handhaben, wird das EM-

Lernverfahren eingesetzt. In jedem Lernschritt des iterativen EM-Lernverfahrens werden basierend auf den aktuellen Parametern Schätzungen (Expected Sufficient Statistics) über die fehlenden Größen erzeugt, die an die Stelle der fehlenden Größen treten. In dem EM-Lernverfahren kann das bedingte Modell $P_B(Z|X,Y)$ dazu verwendet werden, auch für die Variable Z Erwartungswerte oder Expected Sufficient Statistics-Werte zu ermitteln und so dieses Lernverfahren konsistent zu erweitern, um ein gemeinsames Modell verteilter Daten zu erzeugen.

Somit hat die Bank nunmehr die gesamte statistische Information verfügbar und kann entsprechende Analysen über die Daten durchführen.

In diesem Zusammenhang ist anzumerken, dass das oben beschriebene Szenario auch umgekehrt durchgeführt werden kann, d.h. dass die Bank ein statistisches Modell über die zweite Kunden-Datenbank erstellt und dieses an das Autohaus übermittelt, welches seinerseits ein statistisches Gesamt-Modell bildet. Für das Autohaus wäre es beispielsweise wünschenswert, das Alter seiner Kunden zu kennen, deren Familienstand und deren Gehaltseingang, oder jedenfalls eine Schätzung des Alters, des Familienstandes und des Gehaltseingangs. Basierend auf diesen Informationen können den Kunden somit passende Produkte viel gezielter angeboten werden, beispielsweise ist einer jungen Familie mit einem durchschnittlichen Gehaltseingang sicherlich ein anderes Auto anzubieten als einem Single mit einem hohen Gehalt.

Fig.4 zeigt eine Computer-Anordnung 400 gemäß einem vierten Ausführungsbeispiel der Erfindung.

Gemäß diesem Ausführungsbeispiel sind eine Vielzahl von n Computern 401, 413, 420 vorgesehen, die jeweils in

Übereinstimmung mit dem dritten Ausführungsbeispiel eine Kunden-Datenbank führen.

Der erste Computer 401 weist einen Speicher 402, einen Prozessor 403 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 404 auf, welche mittels eines Computerbusses 405 miteinander gekoppelt sind.

Der erste Computer 401 ist ein Computer eines Autohauses, welches in der in dem Speicher 402 gespeicherten Kunden-Datenbank Informationen zu Vorname und Nachname der Kunden, über Wohnort und genutzten Fahrzeugtyp, nicht jedoch über Alter, Familienstand und Gehaltseingang enthält.

Über die Kunden-Datenbank wird von dem ersten Computer 401 ein erstes statistisches Modell 406 gebildet und in dem Speicher 402 gespeichert.

Der zweite Computer 413 weist einen Speicher 414, einen Prozessor 415 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 416 auf, welche mittels eines Computerbusses 417 miteinander gekoppelt sind.

Der zweite Computer 413 ist ein Computer einer Bank, welche in der in dem Speicher 414 gespeicherten Kunden-Datenbank die im dritten Ausführungsbeispiel genannten Informationen enthält. Über die zweite Kunden-Datenbank wird von dem zweiten Computer 413 ein zweites statistisches Modell 418 gebildet und in dem Speicher 414 gespeichert.

Der n-te Computer 420 hat ebenfalls eine Kunden-Datenbank gespeichert. Der n-te Computer 420 weist einen Speicher 421, einen Prozessor 422 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 423 auf, welche mittels eines

Computerbusses 424 miteinander gekoppelt sind. Über die Kunden-Datenbank in dem n-ten Computer 420 ist ebenfalls mittels des EM-Lernverfahrens ein statistisches Modell 425 gebildet und in dem Speicher 421 des n-ten Computers 420 gespeichert.

Die Computer 401, 413, 420 sind mittels einer jeweiligen Kommunikationsverbindung 408 mit einer Client-Computer 409.

Der Client-Computer 409 weist einen Speicher 411, einen Prozessor 412 sowie eine zur Kommunikation gemäß den TCP/IP-Kommunikationsprotokollen eingerichtete Eingangs-/Ausgangsschnittstelle 410 auf, welche mittels eines Computerbusses 426 miteinander gekoppelt sind.

Die Computer 401, 413, 420 übermitteln die statistischen Modelle 406, 418, 525 an den Client-Computer 409 in jeweiligen elektronischen Nachrichten 407, 419, 427, welcher diese in dessen Speicher 410 speichert.

Im Folgenden wird zur einfacheren Darstellung das Ausführungsbeispiel nur unter Berücksichtigung des ersten statistischen Modells 406 und des zweiten statistischen Modells 418 näher erläutert. Es ist jedoch anzumerken, dass erfindungsgemäß eine beliebige Anzahl statistischer Modelle zu einem Gesamt-Modell zusammengeführt werden kann, beispielsweise mittels wiederholten Durchführens der im Folgenden beschriebenen Verfahrensschritte.

Im Unterschied zu dem dritten Ausführungsbeispiel ist es gemäß dem dritten Ausführungsbeispiel das Ziel, mehrere statistische Modelle miteinander zu einem Gesamt-Modell zu kombinieren.

Somit wird in Anlehnung an die im dritten Ausführungsbeispiel verwendeten Nomenklatur von dem Partner A ebenfalls ein statistisches Modell $P_A(W,X,Y)$ erstellt und dann werden die

Modelle $P_A(W,X,Y)$ und $P_B(X,Y,Z)$ zu einem statistischen Gesamt-Modell $P(W,X,Y,Z)$ kombiniert.

Das Gesamt-Modell $P(W,X,Y,Z)$ kann basierend auf den beiden Modellen $P_A(W,X,Y)$ und $P_B(X,Y,Z)$ definiert werden als:

- $P(W,X,Y,Z) = P_A(W,X,Y)P_B(Z|X,Y)$ oder als
- $P(W,X,Y,Z) = P_B(X,Y,Z)P_A(W|X,Y)$.

Auch Kombinationen aus beiden Vorgehensweisen sind erfindungsgemäß vorgesehen. Für den Partner A ist es am sinnvollsten, die erste obige Alternative zu wählen. Damit verfügt er über ein statistisches Gesamt-Modell 426, welches ihm in einer approximativen Weise ermöglicht, auch die Abhängigkeiten zwischen den Attributen W und Z zu analysieren (in diesem Ausführungsbeispiel die Abhängigkeit zwischen Fahrzeugtyp und Gehaltseingang). Basierend auf dem Gesamt-Modell 426 werden beispielsweise bedingte Wahrscheinlichkeitsverteilungen der Form $P(X|Z)$, z.B. eine Verteilung über oder eine Affinität zu Fahrzeugtypen bei einem gegebenen Gehaltseingang, ermittelt. Hierzu wird über die Variablen X und Y marginalisiert.

Zur Erläuterung wird angenommen, dass die Ergebnisse aus dem Gesamt-Modell 426 in einer Art eines zweistufigen Prozesses zustande kommen. Zunächst wird aus der Variable W auf die gemeinsamen Variablen X und Y basierend auf dem Modell $P_A(W,X,Y)$ geschlossen. Entsprechend allen danach erlaubten Kombinationen für die Variablen X und Y wird die bedingte Wahrscheinlichkeitsverteilung $P_B(Z|X,Y)$ (Prädiktion der Variable Z aus den Variablen X und Y) genutzt, um die Verteilung für die Variable Z zu bestimmen.

Im Unterschied zu dem Fall, in dem alle vier Variablen in einer Datenbank zu finden sind, erfolgt die Schlussfolgerung somit erfindungsgemäß indirekt; ähnlich wie bei einer Flüsterpost können dabei Informationen verloren gehen.

Im schlimmsten Fall, nämlich wenn kein Überlapp zwischen den beiden statistischen Abbildern vorliegt, dann ist auch keine Kombination der beiden Modelle möglich. Allerdings ist beispielsweise für den Fall, dass gemeinsame Variablen in den beiden Modellen vorhanden sind, möglich, ein Gesamt-Modell zu bilden, selbst wenn in den beiden Ausgangs-Datenbanken keine gemeinsamen Kunden, beispielsweise kein gemeinsamer Kundenschlüssel, vorhanden ist.

Das Gesamt-Modell 426 P(W,X,Y,Z) kann numerisch einfach gehandhabt werden, wenn der Überlapp zwischen diesen statistischen Modellen nicht zu groß ist, vorzugsweise kleiner als 10 gemeinsame Variablen. In dem Fall eines großen „Überlapp-Raums" können zusätzliche Approximationen verwendet werden, um die Ausführung der folgenden Summen zu beschleunigen, welche gemäß den obigen Ausführungsbeispielen über alle gemeinsamen Zustände der gemeinsamen Variablen X und Y gebildet werden müssen:

$$P(W|Z) \propto \sum_{x,y} P_A(W, X, Y) \cdot P_B(Z|X, Y)$$

bzw.

$$P(W, Z) = \sum_{x,y} P_A(W, X, Y) \cdot P_B(Z|X, Y).$$

Die Summen können insbesondere sehr geschickt approximiert werden basierend auf einem Ansatz durch Einführen einer zusätzlichen künstlichen Variable H und zusätzlichen bedingten Verteilungen (Tafeln im Falle diskreter Variable) $P(H|X, Y)$ und $P(Z|H)$ der Form:

$$P_{approx}(W, Z) \approx \sum_{x,y} P_A(W, X, Y) \sum_{h} P(H \mid X, Y) \cdot P_B(Z \mid H)$$

bzw.

$$P_{approx}(W, X, Y, Z) \approx P_A(W, X, Y) \sum_h P(H \mid X, Y) \cdot P_B(Z \mid H).$$

Die Struktur bzw. die Parametrisierung der bedingten Verteilungen $P(H|X, Y)$ und $P(Z|H)$ bzw. die Form der Abhängigkeit zwischen X,Y und H einerseits und H und Z andererseits wird so gewählt, dass die obigen Summen einfach auszuführen sind. Die Parameter der bedingten Verteilungen $P(H|X, Y)$ und $P(Z|H)$ werden so bestimmt, dass die approximative Gesamtverteilung $P_{approx}(W, X, Y, Z)$ möglicht gut der gewünschten Verteilung

$$P(W, X, Y, Z) = P_A(W, X, Y) \cdot P_B(Z|X, Y)$$

entspricht. Als Kostenfunktion kann hierbei insbesondere die Log-Likelihood bzw. die Kullback-Leibler-Distanz verwendet werden. Als Optimierungsverfahren bieten sich daher wiederum ein EM-Lernverfahren oder ein Gradienten-basiertes Lernverfahren an.

Das Auffinden optimaler Parameter kann und darf durchaus rechenaufwendig sein. Sobald die beiden Wahrscheinlichkeitsmodelle dann zu einem Gesamtmodell „fusioniert" sind kann das Gesamtmodell in einer sehr effizienten Art und Weise genutzt werden.

Es bietet sich insbesondere an, die Variable H als eine versteckte Variable einzuführen, also die Verteilung P(W,X,Y,H) zu parametrisieren als

$$P(W, X, Y, H) = P(H) \cdot P(W, X, Y|H)$$

mit einer so genannten a priori Verteilung P(H).

In dem Fall in dem das Modell P(W,X,Y) bereits ursprünglich als ein Latent Variable Model parametrisiert wurde,

$$P_A(W, X, Y) = \sum_h P_A(X, Y, Z \mid H) \cdot P_A(H),$$

kann unmittelbar die bereits vorhandene latente Variable H genutzt werden.

Statt einer versteckten Variable H können auch mehrere Variablen eingeführt werden. Gleichzeitig kann auch für das Modell PB zur Vereinfachung der Numerik eine versteckte Variable K eingeführt werden. Eine Approximation des Gesamtmodells P(W,X,Y,Z) nimmt damit z.B. die Form an

$$P(W, X, Y, Z) \approx \sum_h P_A(X, Y, Z \mid H) \cdot P_A(H) \sum_k P(K \mid H) \cdot P_B(Z \mid K).$$

In diesem Modell können Summen über den Raum des Überlapps bestehend aus X und Y einfach durch bekannte Inferenzverfahren (beispielsweise das so genannte Junction-Tree-Verfahren) ausgeführt werden. Für die Fusion der beiden Modelle ist lediglich die bedingte Verteilung $P(K|H)$ durch bekannte Lernverfahren zu bestimmen.

Um das Ziel zu erreichen kleine, austauschbare jedoch aber sehr genaue „Abbilder einer Datenbank" zu generieren, sind insbesondere sehr skalierbare Lernverfahren, die hoch komprimierte Abbilder generieren, erwünscht. Gleichzeitig sollen sich die Abbilder effizient fusionieren, d.h. zusammenführen lassen, wozu man insbesondere auch sehr effizient mit fehlenden Informationen umgehen können sollte. Bekannte Lernverfahren sind insbesondere dann langsam, wenn in den Daten viele der Belegungen der Felder fehlen.

Fig.5 zeigt eine Computer-Anordnung 500 gemäß einem fünften Ausführungsbeispiel der Erfindung.

Die Computer-Anordnung 500 wird im Rahmen des Austauschs von Kundeninformation, gemäß diesem Ausführungsbeispiel im Rahmen des Austauschs von Adressinformation von Kunden, eingesetzt. Die Computer-Anordnung 500 weist einen Server-Computer 501 sowie einen oder mehrere mit diesem über ein Telekommunikationsnetz 502 verbundenen Client-Computer 503 auf.

Der Server-Computer 501 weist einen Speicher 504, einen Prozessor 505 sowie eine zur Kommunikation über das Internet eingerichtete Eingangs-/Ausgangsschnittstelle 506 auf, welche Komponenten mittels eines Computerbusses 507 miteinander gekoppelt sind. Der Server-Computer 501 dient gemäß diesem Ausführungsbeispiel als Web-Server-Computer, wie im Folgenden noch näher erläutert wird.

In dem Speicher 504 ist eine große Kunden-Datenbank 508 (insbesondere mit Adressinformation über die Kunden und das Kaufverhalten der Kunden beschreibende Information) gespeichert. Ferner ist in dem Speicher 504 noch ein statistisches Modell 509, welches von dem Server-Computer 501 über die Kunden-Datenbank 508 gebildet worden ist, gespeichert, welches die statistischen Zusammenhänge der in der Kunden-Datenbank 508 enthaltenen Datenelemente repräsentiert.

Das statistische Modell 509 wird unter Verwendung des an sich bekannten EM-Lernverfahrens gebildet. Andere alternative bevorzugt eingesetzte Verfahren zum Bilden des statistischen Modells 509 werden im Folgenden noch im Detail beschrieben.

Gemäß diesem Ausführungsbeispiel der Erfindung wird das statistische Modell 509 automatisch in regelmäßigen vorgegebenen Zeitintervallen erneut, jeweils basierend auf den aktuellsten Daten der Kunden-Datenbank 508, gebildet.

Das statistische Modell 509 wird von dem Server-Computer 501 automatisch zur Übertragung an den oder an mehrere Client-Computer 503 bereitgestellt.

Der Client-Computer 503 weist ebenfalls eine Eingangs-/Ausgangsschnittstelle 510, eingerichtet zur Kommunikation gemäß dem TCP/IP-Kommunikationsprotokoll auf sowie einen Prozessor 511 und einen Speicher 512. Die Komponenten des Client-Computers sind mittels eines Computerbusses 513 miteinander gekoppelt.

Das in einer elektronischen Nachricht 514 von dem Server-Computer 501 an den Client-Computer 503 übertragene statistische Modell 509 wird in dem Speicher 512 des Client-Computers 503 gespeichert.

In diesem Zusammenhang ist anzumerken, dass in dem statistischen Modell 509 die Details der Kunden-Datenbank 508, insbesondere die tatsächlichen Adressen der Kunden, nicht enthalten ist. Das statistische Modell 509 enthält allerdings statistische Information über das Verhalten, insbesondere über das Kaufverhalten der Kunden.

Der Benutzer des Client-Computers 503 wählt nunmehr eine für ihn interessante Gruppe von Kunden, d.h. einen für ihn interessanten Teil 515 des statistischen Modells 509, der ein für das Unternehmen des Benutzers des Client-Computers 503 interessierendes Kaufverhalten beschreibt, aus. Die Information 515 über den ausgewählten Teil des statistischen Modells 509 überträgt der Client-Computer 503 in einer zweiten elektronischen Nachricht 516 zu dem Server-Computer 501.

Unter Verwendung der empfangenen Information liest der Server-Computer 501 die mittels des Teils 515 des statistischen Modells 509 bezeichneten Kunden und die zugehörige Kunden-Detailinformation 517, insbesondere die

Adressen der Kunden, aus der Kunden-Datenbank 508 aus und übermittelt die ausgelesene Kunden-Detailinformation 517 in einer dritten elektronischen Nachricht 518 zu dem Client-Computer 503.

Auf diese Weise ist es möglich, beispielsweise für eine Marketing-Kampagne seitens des Benutzers des Client-Computers 503 gezielt die Adressen der gemäß der Kunden-Datenbank 508 für die Kampagne interessantesten Kunden des Unternehmens des Server-Computers 501 auszuwählen und von dem Server-Computer 501 zu erbitten. Ein erheblicher Vorteil ist ferner darin zu sehen, dass der Server-Computer 501 nur die Informationen an den Client-Computer 503 übermittelt, die auch an diesen übermittelt werden dürfen.

Diese Übermittlung erfolgt gemäß einer Ausgestaltung der Erfindung gegen Bezahlung. Anders ausgedrückt wird somit eine sehr effizientes so genanntes „On-Line Listbroking" realisiert.

Im Folgenden werden verschiedene skalierbare Verfahren zum Bilden eines statistischen Modells angegeben.

Zur besseren Veranschaulichung der bevorzugt eingesetzten Verbesserung eines EM-Lernverfahrens im Falle eines Naiven Bayesianischen Cluster Modells werden im Folgenden einige Grundlagen des EM-Lernverfahrens näher erläutert:

Mit $X = \{X_k, k = 1, \ldots, K\}$ wird einen Satz von K statistischen Variablen (die z.B. den Feldern einer Datenbank entsprechen können) bezeichnet.

Die Zustände der Variablen werden mit kleinen Buchstaben bezeichnet. Die Variable X_1 kann die Zustände $x_{1,1}, x_{1,2}, \ldots$ annehmen, d.h. $X_1 \in \{x_{1,i}, i = 1, \ldots, L_1\}$. L_1 ist die Anzahl der Zustände der Variable X_1. Ein Eintrag in einem Datensatz

(einer Datenbank) besteht nun aus Werten für alle Variablen, wobei $x^{\pi} \equiv \left(x_1^{\pi}, x_2^{\pi}, x_3^{\pi}, \ldots\right)$ den π-ten Datensatz bezeichnet. In dem π-ten Datensatz ist die Variable X_1 in dem Zustand x_1^{π}, die Variable X_2 in dem Zustand x_2^{π}, usw. Die Tafel hat M Einträge, d.h. $\left\{x^{\pi}, \pi = 1, \ldots, M\right\}$. Zusätzlich gibt es eine versteckte Variable oder eine Cluster-Variable, die im Folgenden mit Ω bezeichnet wird; deren Zustände sind $\left\{\omega_i, i = 1, \ldots, N\right\}$. Es gibt also N Cluster.

In einem statistischen Clustering-Modell beschreibt $P(\Omega)$ eine a priori Verteilung; $P(\omega_i)$ ist das a priori Gewicht des i-ten Clusters und $P(\underline{X}|\omega_i)$ beschreibt die Struktur des i-ten Clusters oder die bedingte Verteilung der beobachtbaren (in der Datenbank enthaltenen) Größen $\underline{X} = \left\{X_k, k = 1, \ldots, K\right\}$ in dem i-ten Cluster. Die a priori Verteilung und die bedingten Verteilungen für jedes Cluster parametrisieren zusammen ein gemeinsames Wahrscheinlichkeitsmodell auf $\underline{X} \cup \Omega$ bzw. auf $\underline{X}$.

In einem Naiven Bayesian Network wird vorausgesetzt, dass $p(\underline{X}|\omega_i)$ mit $\prod_{k=1}^{K} p(X_k|\omega_i)$ faktorisiert werden kann.

Im Allgemeinen wird darauf gezielt, die Parameter des Modells, also die a priori Verteilung $p(\Omega)$ und die bedingten Wahrscheinlichkeitstafeln $p(\underline{X}|\omega)$ derart zu bestimmen, dass das gemeinsame Modell die eingetragenen Daten möglichst gut widerspiegelt. Ein entsprechendes EM-Lernverfahren besteht aus einer Reihe von Iterationsschritten, wobei in jedem Iterationsschritt eine Verbesserung des Modells (im Sinne einer so genannten Likelihood) erzielt wird. In jedem Iterationsschritt werden neue Parameter $p^{neu}(...)$ basierend auf den aktuellen oder „alten" Parametern $p^{alt}(...)$ geschätzt.

Jeder EM-Schritt beginnt zunächst mit dem E-Schritt, in dem „Sufficient Statistics" in dafür bereitgehaltenen Tafeln

ermittelt werden. Es wird mit Wahrscheinlichkeitstafeln begonnen, deren Einträge mit Null-Werten initialisiert werden. Die Felder der Tafeln werden im Verlauf des E-Schrittes mit den so genannten Sufficient Statistics $S(\Omega)$ und $S(\underline{X}, \Omega)$ gefüllt, indem für jeden Datenpunkt die fehlenden Informationen (also insbesondere die Zuordnung jedes Datenpunktes zu den Clustern) durch Erwartungswerte ergänzt werden.

Um Erwartungswerte für die Clustervariable Ω zu berechnen ist die a posteriori Verteilung $p^{alt}\left(w_i|\underline{x}^\pi\right)$ zu ermitteln. Dieser Schritt wird auch als „Inferenzschritt" bezeichnet.

Im Falle eines Naive Bayesian Network ist die a posteriori Verteilung für Ω nach der Vorschrift

$$p^{alt}\left(w_i|\underline{x}^\pi\right) = \frac{1}{Z^\pi} p^{alt}(w_i) \prod_{k=1}^{K} p^{alt}\left(x_k^\pi|\omega_i\right)$$

für jeden Datenpunkt $\underline{x}^\pi$ aus den eingetragenen Informationen zu berechnen, wobei $\frac{1}{Z^\pi}$ eine vorgebbare Normierungskonstante ist.

Das Wesentliche dieser Berechnung besteht aus der Bildung des Produkts $p^{alt}\left(x_k^\pi|\omega_i\right)$ über alle $k = 1,\ldots,K$. Dieses Produkt muss in jedem E-Schritt für alle Cluster $i = 1,\ldots,N$ und für alle Datenpunkte $x^\pi, \pi = 1,\ldots,M$ gebildet werden.

Ähnlich aufwendig oft noch aufwendiger ist der Inferenzschritt für die Annahme anderer Abhängigkeitsstrukturen als einem Naive Bayesian Network, und beinhaltet damit den wesentlichen numerischen Aufwand des EM-Lernens.

Die Einträge in den Tafeln $S(\Omega)$ und $S(\underline{X}, \Omega)$ ändern sich nach Bildung des obigen Produktes für jeden Datenpunkt $\underline{x}^{\pi}, \pi = 1, \ldots, M$, da $S(\omega_i)$ um $p^{alt}\left(\omega_i|\underline{x}^{\pi}\right)$ für alle i addiert wird, bzw. eine Summe alle $p^{alt}\left(\omega_i|\underline{x}^{\pi}\right)$ gebildet wird. Auf entsprechende Weise wird $S(\underline{x}, \omega_i)$ (bzw. $S(x_k, \omega_i)$ für alle Variabeln k im Falle eines Naive Bayesian Network) jeweils um $p^{alt}\left(\omega_i|\underline{x}^{\pi}\right)$ für alle Cluster i addiert. Dieses schließt zunächst den E (Expectation)-Schritt ab.

Anhand dieses Schrittes werden neue Parameter $p^{neu}(\Omega)$ und $p^{neu}(\underline{x}|\Omega)$ für das statistische Modell berechnet, wobei $p(\underline{x}|\omega_i)$ die Struktur des i-ten Cluster oder die bedingte Verteilung der in der Datenbank enthaltenden Größen $\underline{X}$ in diesem i-ten Cluster darstellt.

Im M (Maximisation)-Schritt werden unter Optimierung einer allgemeinen log Likelihood

$$L = \sum_{\pi=1}^{M} \log \sum_{i=1}^{N} p\left(\underline{x}^{\pi}|\omega_i\right)p(\omega_i) \qquad (1)$$

neue Parameter $p^{neu}(\Omega)$ und $p^{neu}(\underline{x}|\Omega)$, welche auf den bereits berechneten Sufficient Statistics basieren, gebildet.

Der M-Schritt bringt keinen wesentlichen numerischen Aufwand mehr mit sich.

Somit ist klar, dass der wesentliche Aufwand des Algorithmus in dem Inferenzschritt bzw. auf die Bildung des Produktes $\prod_{k=1}^{K} p^{alt}\left(x_k^{\pi}|\omega_i\right)$ und auf die Akkumulierung der Sufficient Statistics ruht.

Die Bildung von zahlreichen Null-Elementen in den Wahrscheinlichkeitstafeln $p^{alt}(\underline{x}|\omega_i)$ bzw. $p^{alt}(x_k|\omega_i)$ lässt sich jedoch durch geschickte Datenstrukturen und Speicherung von Zwischenergebnissen von einem EM-Schritt zum nächsten dazu ausnutzen, die Produkte effizient zu berechen.

Zum Beschleunigen des EM-Lernverfahrens wird die Bildung eines Gesamtproduktes in einem obigem Inferenzschritt, welcher aus Faktoren von a posteriori Verteilungen von Zugehörigkeitswahrscheinlichkeiten für alle eingegebene Datenpunkte besteht, wie gewöhnlich durchgeführt wird, sobald die erste Null in den dazu gehörenden Faktoren auftritt, wird die Bildung des Gesamtproduktes jedoch abgebrochen. Es lässt sich zeigen, dass für den Fall, dass in einem EM-Lernprozess ein Cluster für einen bestimmten Datenpunkt das Gewicht Null zugeordnet bekommt, dieser Cluster auch in allen weiteren EM-Schritten für diesen Datenpunkt das Gewicht Null zugeordnet bekommen wird.

Somit wird eine sinnvolle Beseitigung von überflüssigen numerischen Aufwand gewährleistet, indem entsprechende Ergebnisse von einem EM-Schritt zum nächsten zwischengespeichert werden und nur für die Cluster, die nicht das Gewicht Null haben, bearbeitet werden.

Es ergeben sich somit die Vorteile, dass aufgrund des Bearbeitungsabbruchs beim Auftreten eines Clusters mit Null Gewichten nicht nur innerhalb eines EM-Schrittes sondern auch für alle weiteren Schritte, besonders bei der Bildung des Produkts im Inferenzschritt, das EM-Lernverfahren insgesamt deutlich beschleunigt wird.

Im Verfahren zur Ermittlung einer in vorgegebenen Daten vorhandenen Wahrscheinlichkeitsverteilung werden Zugehörigkeitswahrscheinlichkeiten zu bestimmten Klassen nur bis zu einem Wert nahezu 0 in einem iterativen Verfahren berechnet, und die Klassen mit

Zugehörigkeitswahrscheinlichkeiten unterhalb eines auswählbaren Wertes im iterativen Verfahren nicht weiter verwendet.

In einer Weiterbildung des Verfahrens wird eine Reihenfolge der zu berechnenden Faktoren derart bestimmt, dass der Faktor, der zu einem selten auftretenden Zustand einer Variabel gehört, als erstes bearbeitet wird. Die selten auftretenden Werte können vor Beginn der Bildung des Produkts derart in einer geordneten Liste gespeichert werden, dass die Variabeln je nach Häufigkeit ihrer Erscheinung einer Null in der Liste geordnet sind.

Es ist weiterhin vorteilhaft, eine logarithmische Darstellung von Wahrscheinlichkeitstafeln zu benutzen.

Es ist weiterhin vorteilhaft, eine dünne Darstellung (sparse representation) der Wahrscheinlichkeitstafeln zu benutzen, z.B. in Form einer Liste, die nur die von Null verschiedenen Elemente enthält.

Ferner werden bei der Berechnung von Sufficient Statistics nur noch die Cluster berücksichtigt, die ein von Null verschiedenes Gewicht haben.

Die Cluster, die ein von Null verschiedenes Gewicht haben, können in eine Liste gespeichert werden, wobei die in der Liste gespeicherte Daten Pointer zu den entsprechenden Cluster sein können.

Das Verfahren kann weiterhin ein Expectation Maximisation Lernprozess sein, bei dem in dem Fall dass für ein Datenpunkt ein Cluster ein a posteriori Gewicht „Null" zugeordnet bekommt, dieser Cluster in allen weiteren Schritten des EM-Verfahrens für diesen Datenpunkt das Gewicht Null erhält und dass dieser Cluster in allen weiteren Schritten nicht mehr berücksichtigt werden muss.

Das Verfahren kann dabei nur noch über Cluster laufen, die ein von Null verschiedenes Gewicht haben.

I. Erstes Beispiel in einem Inferenzschritt

a) Bildung eines Gesamtproduktes mit Unterbrechung bei Nullwert

Für jeden Cluster ω_j in einem Inferenzschritt wird die Bildung eines Gesamtproduktes durchgeführt. Sobald die erste Null in den dazu gehörenden Faktoren, welche beispielsweise aus einem Speicher, Array oder einer Pointerliste herausgelesen werden können, auftritt, wird die Bildung des Gesamtproduktes abgebrochen.

Im Falle des Auftretens eines Nullwertes wird dann das zu dem Cluster gehörende a posteriori Gewicht auf Null gesetzt. Alternativ kann auch zuerst geprüft werden, ob zumindest einer der Faktoren in dem Produkt Null ist. Dabei werden alle Multiplikationen für die Bildung des Gesamtproduktes nur dann durchgeführt, wenn alle Faktoren von Null verschieden sind.

Wenn hingegen bei einem zu dem Gesamtprodukt gehörender Faktor kein Nullwert auftritt, so wird die Bildung des Produktes wie normal fortgeführt und der nächste Faktor aus dem Speicher, Array oder der Pointerliste herausgelesen und zur Bildung des Produktes verwendet.

b) Auswahl einer geeigneten Reihenfolge zur Beschleunigung der Datenverarbeitung

Eine geschickte Reihenfolge wird derart gewählt, dass, falls ein Faktor in dem Produkt Null ist, dieser Faktor mit hoher Wahrscheinlichkeit sehr bald als einer der ersten Faktoren in dem Produkt auftritt. Somit kann die Bildung des Gesamtproduktes sehr bald abgebrochen werden. Die Festlegung

der neuen Reihenfolge kann dabei entsprechend der Häufigkeit, mit der die Zustände der Variablen in den Daten auftreten, erfolgen. Es wird ein Faktor der zu einer sehr selten auftretenden Zustand einer Variable gehört, als erstes bearbeitet. Die Reihenfolge, in der die Faktoren bearbeitet werden, kann somit einmal vor dem Start des Lernverfahrens festgelegt werden, indem die Werte der Variablen in einer entsprechend geordneten Liste gespeichert werden.

c) Logarithmische Darstellung der Tafeln

Um den Rechenaufwand des oben genannten Verfahrens möglichst einzuschränken, wird vorzugsweise eine logarithmische Darstellung der Tafeln benutzt, um beispielsweise Underflow-Probleme zu vermeiden. Mit dieser Funktion können ursprünglich Null-Elemente zum Beispiel durch einen positiven Wert ersetzt werden. Somit ist eine aufwendige Verarbeitung bzw. Trennungen von Werten, die nahezu Null sind und sich voneinander durch einen sehr geringen Abstand unterscheiden, nicht weiter notwendig.

d) Umgehung von erhöhter Summierung bei der Berechnung von Sufficient Statistics

In dem Fall, dass die dem Lernverfahren zugegebenen stochastischen Variablen eine geringe Zugehörigkeitswahrscheinlichkeit zu einem bestimmten Cluster besitzen, werden im Laufe des Lernverfahrens viele Cluster das a posteriori Gewicht Null haben.

Um auch das Akkumulieren der Sufficient Statistics in dem darauf folgenden Schritt zu beschleunigen, werden nur noch solche Cluster in diesem Schritt berücksichtigt, die ein von Null verschiedenes Gewicht haben.

Dabei ist es vorteilhaft, die von Null verschiedenen Cluster in einer Liste, einem Array oder einer ähnlichen

Datenstruktur gespeichert werden, die es erlaubt, nur die von Null verschiedenen Elemente zu speichern.

II. Zweites Beispiel in einem EM Lernverfahren

a) Nicht-Berücksichtigung von Cluster mit Null-Zuordnungen für einen Datenpunkt

Insbesondere wird hier in einem EM-Lernverfahren von einem Schritt des Lernverfahrens zum nächsten Schritt für jeden Datenpunkt gespeichert, welche Cluster durch Auftreten von Nullen in den Tafeln noch erlaubt sind und welche nicht mehr.

Wo im ersten Beispiel Cluster, die durch Multiplikation mit Null ein a posteriori Gewicht Null erhalten, aus allen weiteren Berechnungen ausgeschlossen werden, um dadurch numerischen Aufwand zu sparen, werden in gemäß diesem Beispiel auch von einem EM-Schritt zum nächsten Zwischenergebnisse bezüglich Cluster-Zugehörigkeiten einzelner Datenpunkte (welche Cluster bereits ausgeschlossen bzw. noch zulässig sind) in zusätzlich notwendigen Datenstrukturen gespeichert.

b) Speichern einer Liste mit Referenzen auf relevante Cluster

Für jeden Datenpunkt oder für jede eingegebene stochastische Variable kann zunächst eine Liste oder eine ähnliche Datenstruktur gespeichert werden, die Referenzen auf die relevanten Cluster enthalten, die für diesen Datenpunkt ein von Null verschiedenes Gewicht bekommen haben.

Insgesamt werden in diesem Beispiel nur noch die erlaubten Cluster, allerdings für jeden Datenpunkt in einem Datensatz, gespeichert.

Die beiden obigen Beispiele können miteinander kombiniert werden, was den Abbruch bei „Null"-Gewichten im

Inferenzschritt ermöglicht, wobei in folgenden EM-Schritten nur noch die zulässigen Cluster nach dem zweiten Beispiel berücksichtigt werden.

Eine zweite Variante des EM-Lernverfahrens wird im Folgenden näher erläutert. Es ist darauf hinzuweisen, dass dieses Verfahren unabhängig von der Verwendung des auf diese Weise gebildeten statistischen Modells ist.

Bezugnehmend auf das oben beschriebene EM-Lernverfahren lässt sich zeigen, dass das Ergänzen fehlender Information nicht für alle Größen erfolgen muss. Erfindungsgemäß wurde erkannt, dass ein Teil der fehlenden Information „ignoriert" werden kann. Anders ausgedrückt bedeutet dies, dass nicht versucht wird, etwas über eine Zufallsvariable Y zu lernen aus Daten, in denen keine Information über die Zufallsvariable Y (einem Knoten Y) enthalten ist oder dass nicht versucht wird, etwas über die Zusammenhänge zwischen zwei Zufallsvariablen Y und X (zwei Knoten Y und X) aus Daten, in denen keine Information über die Zufallsvariablen Y und X enthalten ist.

Damit wird nicht nur der numerische Aufwand zur Durchführung des EM-Lernverfahrens wesentlich reduziert, sondern es wird ferner erreicht, dass das EM-Lernverfahren schneller konvergiert. Ein zusätzlicher Vorteil ist darin zu sehen, dass statistische Modelle mittels dieser Vorgehensweise leichter dynamisch aufbauen lassen, d.h. während des Lernprozesses können leichter Variablen (Knoten) in einem Netz, dem gerichteten Graphen, ergänzt werden.

Als anschauliches Beispiel für das erfindungsgemäße Verfahren wird angenommen, dass ein statistisches Modell Variablen enthält, die beschreiben, welche Bewertung ein Kinobesucher einem Film gegeben hat. Für jeden Film gibt es eine Variable, wobei jeder Variable eine Mehrzahl von Zuständen zugeordnet ist, wobei jeder Zustand jeweils einen Bewertungswert repräsentiert. Für jeden Kunden gibt es einen Datensatz, in

dem gespeichert ist, welcher Film welchen Bewertungswert erhalten hat. Wird ein neuer Film angeboten, so fehlen anfangs die Bewertungswerte für diesen Film. Mittels der neuen Variante des EM-Lernverfahrens ergibt sich nunmehr die Möglichkeit, das EM-Lernverfahren bis zu dem Erscheinen des neuen Films nur mit den bis dorthin bekannten Filmen durchzuführen, d.h. den neuen Film (d.h. allgemein den neuen Knoten in dem gerichteten Graphen) zunächst zu ignorieren. Erst mit Erscheinen des neuen Films wird das statistische Modell um eine neue Variable (einen neuen Knoten) dynamisch ergänzt und die Bewertungen des neuen Films werden berücksichtigt. Die Konvergenz des Verfahrens im Sinne der log Likelihood ist dabei noch immer gewährleistet; das Verfahren konvergiert sogar schneller.

Im Folgenden wird erläutert, unter welchen Bedingungen fehlende Informationen nicht berücksichtigt werden müssen.

Zur Erläuterung der Vorgehensweise wird folgende Notation verwendet. Mit H wird ein versteckter Knoten bezeichnet. Mit $\underline{O} = \{O^1, O^2, \ldots, O^M\}$ wird ein Satz von M beobachtbaren Knoten in dem gerichteten Graphen des statistischen Modells bezeichnet.

Es wird ohne Einschränkung der Allgemeingültigkeit im Folgenden ein Bayesianisches Wahrscheinlichkeitsmodell angenommen, welches gemäß folgender Vorschrift faktorisiert werden kann:

$$P(H, \underline{O}) = P(H) \prod_{\pi=1}^{M} P\left(O^{\pi} | H\right). \qquad (2)$$

Es ist in diesem Zusammenhang anzumerken, dass die beschriebene Vorgehensweise auf jedes statistische Modell anwendbar ist, und nicht auf ein Bayesianisches Wahrscheinlichkeitsmodell beschränkt ist, wie später noch im Detail dargelegt wird.

Mit Großbuchstaben werden im Weiteren Zufallsvariablen bezeichnet, wohingegen mit einem Kleinbuchstaben eine Instanz einer jeweiligen Zufallsvariable bezeichnet wird.

Es wird ein Datensatz mit N Datensatzelementen $\{\underline{o}_i, i = 1, \ldots, N\}$ angenommen, wobei für jedes Datensatzelement nur ein Teil der beobachtbaren Knoten tatsächlich beobachtet wird. Für das i-te Datensatzelement wird angenommen, dass die Knoten $\underline{X}_i$ beobachtet wird und dass die Beobachtungswerte der Knoten $\underline{Y}_i$ fehlen.

Es gilt also:

$$\underline{X}_i \cup \underline{Y}_i = \underline{o}_i . \qquad (3)$$

Es ist zu bemerken, dass für jedes Datensatzelement ein unterschiedlicher Satz von Knoten $\underline{X}_i$ beobachtet werden kann, d.h. dass gilt:

$$\underline{X}_i \neq \underline{X}_j \text{ für } i \neq j. \qquad (4)$$

Die Indizes für vorhandene Knoten werden mit κ bezeichnet, d.h. $\underline{X}_i = \{X_i^\kappa, \kappa = 1, \ldots, K_i\}$, die Indizes für nicht vorhandene Knoten werden mit λ bezeichnet, d.h. $\underline{Y}_i = \{Y_i^\lambda, \lambda = 1, \ldots, L_i\}$.

Im Falle eines Bayesianischen Netzes weist das übliche EM-Lernverfahren die folgenden Schritten auf, wie oben schon kurz dargestellt:

1) E-Schritt

Das Verfahren wird mit „leeren" Tabellen $SS(H)$ und $SS(o^\pi, H), i = 1, \ldots, M$ (initialisiert mit „Nullen" gestartet, um darauf basierend die Schätzungen (Sufficient Statistics-Werte) zu akkumulieren. Für jedes Datensatzelement $\underline{o}_i$ werden

die a posteriori Verteilung $P(H|\underline{x}_i)$ für den versteckten Knoten H sowie die a posteriori Verbund-Verteilung $P(H, Y_i^{\pi}|\underline{x}_i)$ für jeden der nicht vorhandenen Knoten $\underline{Y}_i$ zusammen mit dem versteckten Knoten H berechnet.

Für jedes Datensatzelement i werden die Schätzungen für das statistische Modell akkumuliert gemäß folgenden Vorschriften:

$$SS(H) \quad += \quad \sum_i P(H|\underline{x}_i), \tag{5}$$

$$SS(X_i^{\kappa} = x_i^{\kappa}, H) \quad += \quad P(H|\underline{x}_i), \qquad \forall \text{ vorhandenen Knoten } X_i^{\kappa}, \tag{6}$$

$$SS(Y_i^{\lambda}, H) \quad += \quad P(H, Y_i^{\lambda}|\underline{x}_i), \quad \forall \text{ nicht vorhandenen Knoten } Y_i^{\lambda}. \tag{7}$$

Mit dem Symbol += wird die Aktualisierung, d.h. die Akkumulation der Tabellen für die Schätzungen gemäß den Werten der jeweiligen „rechten Seite" der Gleichung bezeichnet.

2) M-Schritt

In dem M-Schritt werden die Parameter für alle Knoten gemäß folgenden Vorschriften aktualisiert:

$$P(H) \propto SS(H), \tag{8}$$

$$P(O^{\pi}|H) \propto SS(O^{\pi}, H), \tag{9}$$

wobei mit dem Symbol $\propto$ angegeben wird, dass die Wahrscheinlichkeits-Tabellen beim Übertragen von SS auf P zu normieren sind.

Gemäß dem EM-Lernverfahren werden die Erwartungswerte für die nicht vorhandenen Knoten $\underline{Y}_i$ berechnet und entsprechend den

Sufficient Statistics-Werten für diese Knoten gemäß Vorschrift (7) aktualisiert.

Andererseits ist das Berechnen und Aktualisieren der Verbund-Verteilung $P(H, Y_i^\lambda | \underline{x}_i)$ für alle Knoten $Y_i^\lambda \in \underline{Y}_i$ sehr rechenaufwendig. Ferner ist das Aktualisieren der Verbund-Verteilung $P(H, Y_i^\lambda | \underline{x}_i)$ ein Grund für das langsame Konvergieren des EM-Lernverfahrens, wenn ein großer Teil an Information fehlt.

Angenommen, die Tabellen werden mit Zufallszahlen initialisiert, bevor das EM-Lernverfahren gestartet wird.

In diesem Fall entspricht die Verbund-Verteilung $P(H, Y_i^\lambda | \underline{x}_i)$ im Wesentlichen diesen Zufallszahlen im ersten Schritt. Dies bedeutet, dass die initialen Zufallszahlen in den Sufficient Statistics-Werten berücksichtigt werden gemäß dem Verhältnis der fehlenden Information bezogen auf die vorhandenen Information. Dies bedeutet, dass die initialen Zufallszahlen in jeder Tabelle nur gemäß dem Verhältnis der fehlenden Information bezogen auf die vorhandenen Information „gelöscht" werden.

Im Folgenden wird bewiesen, dass für den Fall eines Bayesianischen Netzes als statistisches Modell der Schritt gemäß Vorschrift (7) nicht notwendig ist und somit weggelassen bzw. übersprungen werden kann.

Die Log-Likelihood des Bayesianischen Netzes als statistisches Modell ist gegeben durch:

$$L[P] = \sum_{i=1}^{N} \log P(\underline{x}_i). \qquad (10)$$

Für frei vorgegebene Tabellen $B(H|\underline{x}_i)$, welche hinsichtlich dem Knoten H normiert sind, ergibt sich für die Log-Likelihood:

$$\begin{aligned} L[P] &= \sum_{i=1}^{N} B(h|\underline{x}_i) \log P(\underline{x}_i) \\ &= \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log \frac{P(\underline{x}_i, h)}{P(h|\underline{x}_i)} \\ &= \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(\underline{x}_i, h) - \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(h|\underline{x}_i) \end{aligned} \tag{11}$$

Die Summe $\sum_{h}$ bezeichnet die Summe über alle Zustände h des Knotens H.

Unter Verwendung der folgenden Definitionen für R[P, B] und H[P, B]:

$$R[P, B] = \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(\underline{x}_i, h) \tag{12}$$

$$H[P, B] = \sum_{i=1}^{N} \sum_{h} B(h|\underline{x}_i) \log P(h|\underline{x}_i) \tag{13}$$

ergibt sich für die Log-Likelihood gemäß Vorschrift (11):

$$L[P] = R[P, B] - H[P, B]. \tag{14}$$

Allgemein gilt:

$$H[P, B] \leq H[P, P], \tag{15}$$

da $H[P, P] - H[P, B]$ die nicht-negative Kreuzentropie zwischen $P(h|\underline{x}_i)$ und $B(h|\underline{x}_i)$ darstellt.

In dem t-ten Schritt wird das aktuelle statistische Modell mit $P^{(t)}$ bezeichnet. Ausgehend von dem aktuellen statistischen Modell $P^{(t)}$ des t-ten Schrittes wird ein neues statistisches Modell $P^{(t+1)}$ konstruiert derart, dass gilt:

$$R\left[P^{(t+1)}, P^{(t)}\right] > R\left[P^{(t)}, P^{(t)}\right]. \tag{16}$$

Es gilt:

$$\begin{aligned} L\left[P^{(t+1)}\right] &= R\left[P^{(t+1)}, B\right] - H\left[P^{(t+1)}, B\right] \\ &= R\left[P^{(t+1)}, P^{(t)}\right] - H\left[P^{(t+1)}, P^{(t)}\right] \\ &> R\left[P^{(t)}, P^{(t)}\right] - H\left[P^{(t)}, P^{(t)}\right] \\ &= L\left[P^{(t)}\right] \end{aligned} \tag{17}$$

Die erste Zeile gilt allgemein für alle B (vergleiche Vorschrift (14)). Die zweite Zeile der Vorschrift (17) insbesondere für den Fall, dass gilt:

$$B = P^{(t)}. \tag{18}$$

Die dritte Zeile gilt aufgrund Vorschrift (15). Die letzte Zeile von Vorschrift (17) entspricht wiederum Vorschrift (14).

Somit ergibt sich, dass für den Fall $R\left[P^{(t+1)}, P^{(t)}\right] > R\left[P^{(t)}, P^{(t)}\right]$ sicher gilt:

$$L\left[P^{(t+1)}\right] > L\left[P^{(t)}\right]. \tag{19}$$

Es ist auf den Unterschied zu dem Standard-EM-Lernverfahren hinzuweisen [2], bei dem der R-Term definiert ist gemäß folgender Vorschrift:

$$R^{Standard}[P, B] = \sum_{i=1}^{N} \sum_{h, \underline{y}_i} B(\underline{y}_i, h|\underline{x}_i) \log P(\underline{x}_i, \underline{y}_i, h). \quad (20)$$

Es ist anzumerken, dass in dem Argument von P und B in der obigen Vorschrift (20) im Unterschied zu der Definition entsprechend den Vorschriften (12) und (13) auch die fehlenden Größen y auftreten.

Eine Sequenz von EM-Iterationen wird gebildet derart, dass gilt:

$$R^{Standard}[P^{(t+1)}, P^{(t)}] > R^{Standard}[P^{(t)}, P^{(t)}]. \quad (21)$$

Bei dem erfindungsgemäßen Lernverfahren wird für den Fall eines Bayesianischen Netzes eine Sequenz von EM-Iterationen derart gebildet, dass gilt:

$$R[P^{(t+1)}, P^{(t)}] > R[P^{(t)}, P^{(t)}]. \quad (16)$$

Nun wird gezeigt, dass die auf R, definiert gemäß Vorschrift (12), zu dem oben beschriebenen Lernverfahren führt, bei dem Vorschrift (7) übersprungen wird. Bei einem gegebenen aktuellen statistischen Modell $P^{(t)}$ zu einer Iteration t ist es das Ziel des Verfahrens, ein neues statistisches Modell $P^{(t+1)}$ in der Iteration t+1 zu berechnen, indem $R[P, P^{(t)}]$ bezüglich P optimiert wird. Unter Verwendung der Faktorisierung gemäß Vorschrift (2) ergibt sich:

$$R[P, P^{(t)}] = \sum_{i=1}^{N} \sum_{h} P^{(t)}(h|\underline{x}_i) \log P(h) + \sum_{i=1}^{N} \sum_{h} \sum_{\kappa=1}^{K_i} P^{(t)}(h|\underline{x}_i) \log P(x_i^{\kappa}|h). \quad (22)$$

Eine Optimierung von R in Bezug auf das Modell P führt zu dem erfindungsgemäßen Verfahren. Der erste Term führt zu der

Standard-Aktualisierung der P(H) gemäß den Vorschriften (5) und (7).

Mit

$$SS(h) \equiv \sum_{i=1}^{N} P^{(t)}(h|\underline{x}_i) \log P(h) \tag{23}$$

ergibt sich der erste Term von Vorschrift (22) zu

$$\sum_{h} \sum_{i=1}^{N} P^{(t)}(h|\underline{x}_i) \log P(h) = \sum_{h} SS(h) \log P(h), \tag{24}$$

was im Wesentlichen der Kreuzentropie zwischen SS(H) und P(H) entspricht. Somit ist das optimale P(H) durch SS(H) gegeben. Dies entspricht dem M-Schritt gemäß Vorschrift (8).

Der zweite Term von Vorschrift (22) führt zu einer EM-Aktualisierung für die Tabellen der bedingten Wahrscheinlichkeiten $P(o^{\pi}|H)$, wie mittels der Vorschriften (6) und (9) beschrieben. Um dies zu veranschaulichen werden alle die Terme in R gesammelt, welche abhängig sind von $P(o^{\pi}|H)$. Diese Terme sind gegeben gemäß folgender Vorschrift:

$$\sum_{h} \sum_{\substack{i=1 \\ o^{\pi} \in \underline{X}_i}}^{N} P^{(t)}(h|\underline{x}_i) \log P(o^{\pi}|h). \tag{25}$$

Die Summe $\sum_{\substack{i=1 \\ o^{\pi} \in \underline{X}_i}}^{N}$ bezeichnet die Summe über alle Datenelemente i in dem Datensatz, wobei o^{π} einer der beobachteten Knoten ist, d.h. bei dem gilt:

$$O^{\pi} \in \underline{X}_i . \qquad (26)$$

Zusammenfassend kann der obige Ausdruck (25) als die Kreuzentropie zwischen $P\left(O^{\pi}|H\right)$ und den Sufficient Statistics-Werten, welche gemäß Vorschrift (6) akkumuliert werden, interpretiert werden. Es ist somit nicht erforderlich, eine Aktualisierung gemäß Vorschrift (7) vorzusehen. Dies ist auf die Summe $\sum_{\substack{i=1 \\ O^{\pi} \in \underline{X}_i}}^{N}$ in Vorschrift (25) bzw. auf die Summe $\sum_{\kappa=1}^{K_i}$ in Vorschrift (22) zurückzuführen. Diese Summe berücksichtigt nur die beobachteten Knoten, im Gegensatz zu der Definition von $R^{Standard}$ gemäß Vorschrift (20), in der auch die nicht beobachteten Knoten $\underline{Y}_i$ berücksichtigt werden.

Im Folgenden wird in einem allgemeingültigeren Fall die Gültigkeit der Vorgehensweise, nicht beobachtete Knoten im Rahmen der Aktualisierung der Sufficient Statistics Tafeln nicht zu berücksichtigen, dargelegt, womit gezeigt wird, dass die Vorgehensweise nicht auf ein so genanntes Bayesianisches Netz beschränkt ist.

Es wird ein Satz von Variablen $\underline{Z} = \left\{Z^1, Z^2, \ldots, Z^M\right\}$ angenommen. Es wird ferner angenommen, dass das statistische Modell auf folgende Weise faktorisierbar ist:

$$P(\underline{Z}) = \prod_{\sigma=1}^{M} P\left(Z^{\sigma} \middle| \prod \left[Z^{\sigma}\right]\right), \qquad (27)$$

wobei mit $\prod \left[Z^{\sigma}\right]$ die „Eltern"-Knoten des Knoten Z^{σ} in dem Bayesianischen Netz bezeichnet werden. Ferner wird für jeden Knoten $\underline{Z}$ ein Datensatz $\left\{z_i, i = 1, \ldots, N\right\}$ mit N Datensatzelementen angenommen. Wie schon oben angenommen, wird auch in diesem Fall in jedem der N Datensatzelemente ein nur ein Teil der Knoten $\underline{Z}$ beobachtet. Für das i-te

Datensatzelement wird angenommen, dass die Knoten $\underline{X}_i$ beobachtet werden; die Knoten $\overline{X}_i$ werden nicht beobachtet und es gilt:

$$\underline{Z} = \underline{X}_i \cup \overline{X}_i . \tag{28}$$

Für jedes der N Datensatzelemente werden die nicht beobachteten Knoten $\overline{X}_i$ in zwei Untermengen $\underline{H}_i$ und $\underline{Y}_i$ aufgeteilt derart, dass keiner der Knoten in den Mengen $\underline{X}_i$ und $\underline{H}_i$ ein abhängiger, d.h. nachfolgender Knoten („Kinder"-Knoten) eines Knotens in der Menge $\underline{Y}_i$ ist. Anschaulich bedeutet das, dass $\underline{Y}_i$ einem Zweig in einem Bayesianischen Netz entspricht, zu dem es keine Informationen in den Daten gibt.

Somit ergeben sich die Verbund-Verteilungen für die Knoten $\underline{X}_i$ und $\underline{H}_i$ gemäß folgender Vorschrift:

$$P(\underline{X}_i, \underline{H}_i) = \prod_{X \in \underline{X}_i} P\left(X \middle| \prod [X]\right) \prod_{H \in \underline{H}_i} P\left(H \middle| \prod [H]\right). \tag{29}$$

1) E-Schritt

Für jeden Knoten Z werden mit Null-Werten initialisierte Tabellen $SS\left(Z, \prod [Z]\right)$ gebildet bzw. bereitgestellt. Für jedes Datensatzelement i in dem Datensatz werden die a posteriori Verteilung $P\left(Z, \prod [Z] \middle| \underline{X}_i = \underline{x}_i\right)$ berechnet und die Sufficient Statistics-Werte gemäß folgender Vorschrift akkumuliert für jeden Knoten $Z \in \underline{X}_i$ und $Z \in \underline{H}_i$:

$$SS\left(Z, \prod |Z\right) \quad += \quad P\left(Z, \prod [Z] \middle| \underline{X}_i = \underline{x}_i\right). \tag{30}$$

Die Sufficient Statistics-Werte der Tabellen, welche den Knoten in $\overline{X}_i$ zugeordnet sind, werden nicht aktualisiert.

2) M-Schritt

Die Parameter (Tabellen) aller Knoten werden gemäß folgender Vorschrift aktualisiert:

$$P\left(z^{\sigma} \middle| \prod\left[z^{\sigma}\right]\right) \propto SS\left(z^{\sigma}, \prod\left[z^{\sigma}\right]\right). \qquad (31)$$

Anschaulich kann die Erfindung darin gesehen werden, dass ein breiter und einfacher (im Allgemeinen jedoch allerdings approximativer) Zugang zu der Statistik einer Datenbank (bevorzugt über das Internet) durch Bildung statistischer Modelle für die Inhalte der Datenbank geschaffen wird. Somit werden die statistischen Modelle zur „Remote Diagnose", zur so genannten „Remote Assistance" oder zum „Remote Research" über ein Kommunikationsnetz automatisch versendet. Anders ausgedrückt wird „Wissen" in Form eines statistischen Modells kommuniziert und versendet. Wissen ist häufig Wissen über die Zusammenhänge und wechselseitigen Abhängigkeiten in einer Domäne, beispielsweise über die Abhängigkeiten in einem Prozess. Ein statistisches Modell einer Domäne, welches aus den Daten der Datenbank gebildet wird, ist ein Abbild all dieser Zusammenhänge. Technisch stellen die Modelle eine gemeinsame Wahrscheinlichkeitsverteilung der Dimensionen der Datenbank dar, sind also nicht auf eine spezielle Aufgabenstellung eingeschränkt, sondern stellen beliebige Abhängigkeiten zwischen den Dimensionen dar. Komprimiert zu dem statistischen Modell lässt sich das Wissen über eine Domäne sehr einfach handhaben, versenden, beliebigen Nutzern bereitstellen, etc.

Die Auflösung des Abbildes bzw. des statistischen Modells kann entsprechend den Anforderungen des Datenschutzes oder den Bedürfnissen der Partner gewählt werden.

In diesem Dokumenten sind folgende Veröffentlichungen zitiert:

[1] Christopher M. Bishop, Latent Variable Models, M.I. Jordan (Editor), Learning in Graphical Models, Kulwer, 1998, Seiten 371 - 405

[2] M.A. Tanner, Tools for Statistical Inference, Springer, New York, 3. Auflage, 1996, Seiten 64 - 135

[3] Radford M. Neal und Geoffrey E. Hinton, A View of the EM Algorithm that Justifies Incremental, Sparse and Other Variants, M.I. Jordan (Editor), Learning in Graphical Models, Kulwer, 1998, Seiten 355 - 371

[4] D. Heckermann, Bayesian Networks for Data Mining, Data Mining and Knowledge Discovery, Seiten 79 - 119, 1997

[5] Reimar Hofmann, Lernen der Struktur nichtlinearer Abhängigkeiten mit graphischen Modellen, Dissertation an der Technischen Universität München, Verlag: dissertation.de, ISBN:3-89825-131-4

Patentansprüche

1. Verfahren zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank,

- bei dem für die erste Datenbank ein erstes statistisches Modell gebildet wird, welches die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente repräsentiert,
- bei dem das erste statistische Modell in einem Server-Computer gespeichert wird,
- bei dem das erste statistische Modell von dem Server-Computer über ein Kommunikationsnetz zu einem Client-Computer übertragen wird,
- bei dem das empfangene erste statistische Modell von dem Client-Computer weiterverarbeitet wird.

2. Verfahren gemäß Anspruch 1,
bei dem unter Verwendung des ersten statistischen Modells und Datenelementen einer in dem Client-Computer gespeicherten zweiten Datenbank ein statistisches Gesamt-Modell gebildet wird, welches zumindest einen Teil der in dem ersten statistischen Modell und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

3. Verfahren gemäß Anspruch 1,

- bei dem für eine zweite Datenbank ein zweites statistisches Modell gebildet wird, welches die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert,
- bei dem das zweite statistische Modell über das Kommunikationsnetz zu dem Client-Computer übertragen wird ,
- bei dem unter Verwendung des ersten statistischen Modells und des zweiten statistischen Modells von dem Client-Computer ein statistisches Gesamt-Modell gebildet wird, welches zumindest einen Teil der in dem ersten

statistischen Modell und in dem zweiten statistischen Modell enthaltenen statistischen Information aufweist.

4. Verfahren gemäß Anspruch 3,

- bei dem das zweite statistische Modell in einem zweiten Server-Computer gespeichert wird,
- bei dem das zweite statistische Modell von dem zweiten Server-Computer über ein Kommunikationsnetz zu dem Client-Computer übertragen wird.

5. Verfahren gemäß einem der Ansprüche 1 bis 4, bei dem mindestens eines der statistischen Modelle mittels eines skalierbaren Verfahrens gebildet wird, mit dem der Kompressionsgrad des statistischen Modells verglichen mit den in der jeweiligen Datenbank enthaltenen Datenelementen einstellbar ist.

6. Verfahren gemäß einem der Ansprüche 1 bis 5, bei dem mindestens eines der statistischen Modelle mittels eines EM-Lernverfahrens oder mittels eines gradientenbasierten Lernverfahrens gebildet wird.

7. Verfahren gemäß einem der Ansprüche 1 bis 6, bei dem die erste Datenbank oder/und die zweite Datenbank Datenelemente aufweist/aufweisen, welche mindestens eine technische Anlage beschreiben.

8. Verfahren gemäß Anspruch 7, bei dem die die mindestens eine technische Anlage beschreibenden Datenelemente zumindest teilweise an der technischen Anlage gemessene Werte darstellen, welche das Betriebsverhalten der technischen Anlage beschreiben.

9. Verfahren zum rechnergestützten Bilden eines statistischen Modells einer Datenbank, welche eine Vielzahl von Datenelementen aufweist,

- bei dem ein EM-Lernverfahren auf die Datenelemente durchgeführt wird, so dass zu einem vorgebbaren gerichteten Graph statistische Zusammenhänge zwischen den Datenelementen ermittelt werden,
- wobei der gerichtete Graph Knoten und Kanten aufweist,
- wobei die Knoten vorgebbare beobachtbare Datenbank-Zustände und nicht beobachtbare Datenbank-Zustände beschreiben,
- bei dem im Rahmen des EM-Lernverfahrens nur die Erwartungswerte ermittelt werden zu den beobachtbaren Datenbank-Zuständen sowie zu den nicht beobachtbaren Datenbank-Zuständen, deren Eltern-Datenbank-Zustände beobachtbare Datenbank-Zustände sind.

10. Computer-Anordnung zum rechnergestützten Bereitstellen von Datenbankinformation einer ersten Datenbank,

- mit einem Server-Computer, in dem ein erstes statistisches Modell, welches für eine erste Datenbank gebildet ist, gespeichert ist, wobei das erste statistische Modell die statistischen Zusammenhänge der in der ersten Datenbank enthaltenen Datenelemente repräsentiert,
- mit einem mit dem Server-Computer mittels eines Kommunikationsnetz gekoppelten Client-Computer, der eingerichtet ist zur Weiterverarbeitung des von dem Server-Computer über das Kommunikationsnetz zu dem Client-Computer übertragenen ersten statistischen Modells.

11. Computer-Anordnung gemäß Anspruch 10,

- bei der in dem Client-Computer eine zweite Datenbank mit Datenelementen gespeichert ist,
- wobei der Client-Computer eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und den Datenelementen der zweiten Datenbank, aufweist, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten

statistischen Modell und in der zweiten Datenbank enthaltenen statistischen Information aufweist.

12. Computer-Anordnung gemäß Anspruch 10,

- mit einem zweiten Server-Computer, in dem ein zweites statistisches Modell, welches für eine zweite Datenbank gebildet ist, gespeichert ist, wobei das zweite statistische Modell die statistischen Zusammenhänge der in der zweiten Datenbank enthaltenen Datenelemente repräsentiert,
- wobei der Client-Computer mittels des Kommunikationsnetzes mit dem zweiten Server-Computer gekoppelt ist,
- wobei der Client-Computer eine Einheit zum Bilden eines statistischen Gesamt-Modells unter Verwendung des ersten statistischen Modells und des zweiten statistischen Modells, aufweist, wobei das statistische Gesamt-Modell zumindest einen Teil der in dem ersten statistischen Modell und in dem zweiten statistischen Modell enthaltenen statistischen Information aufweist.

FIG 1

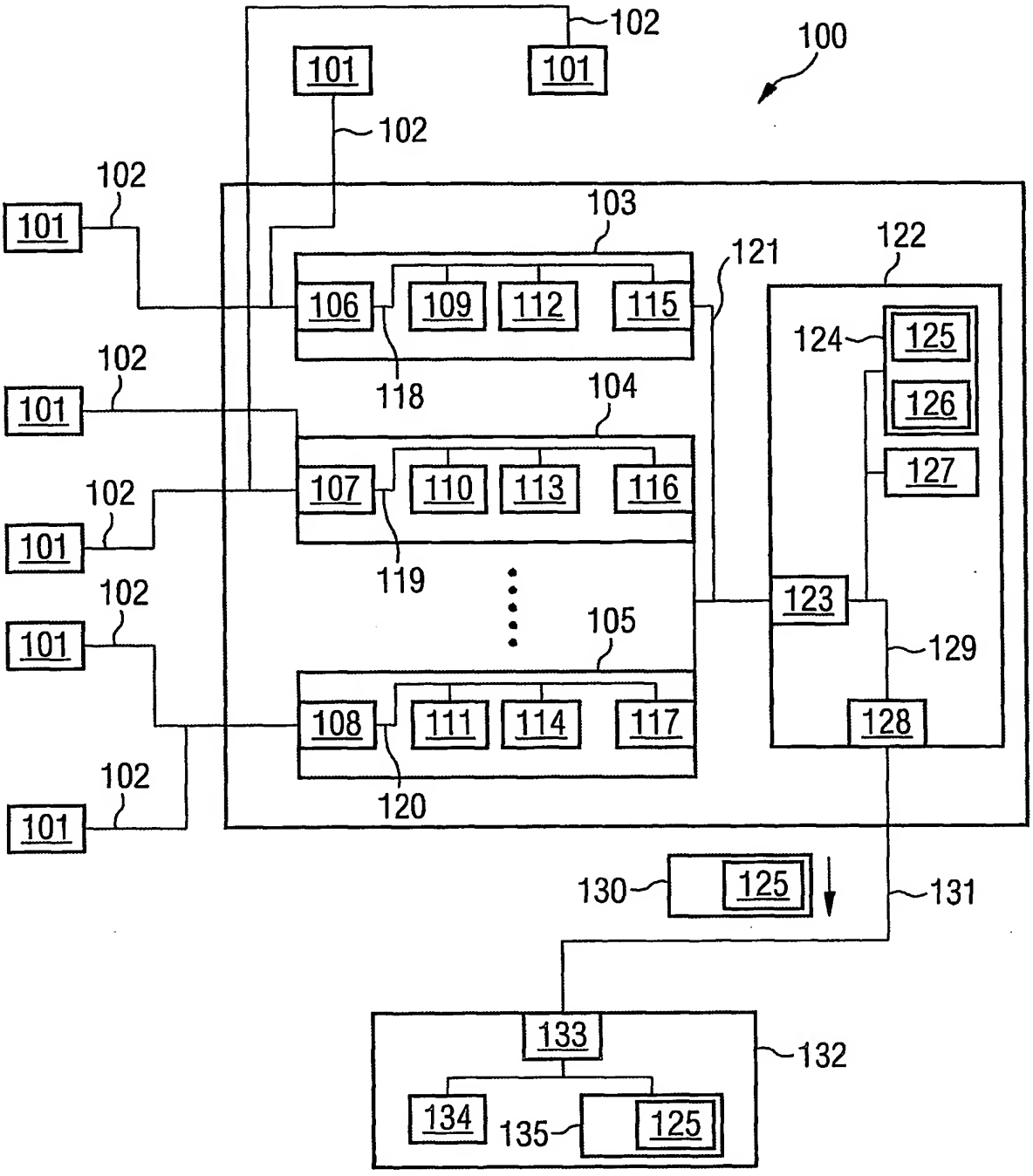

100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135

FIG 2

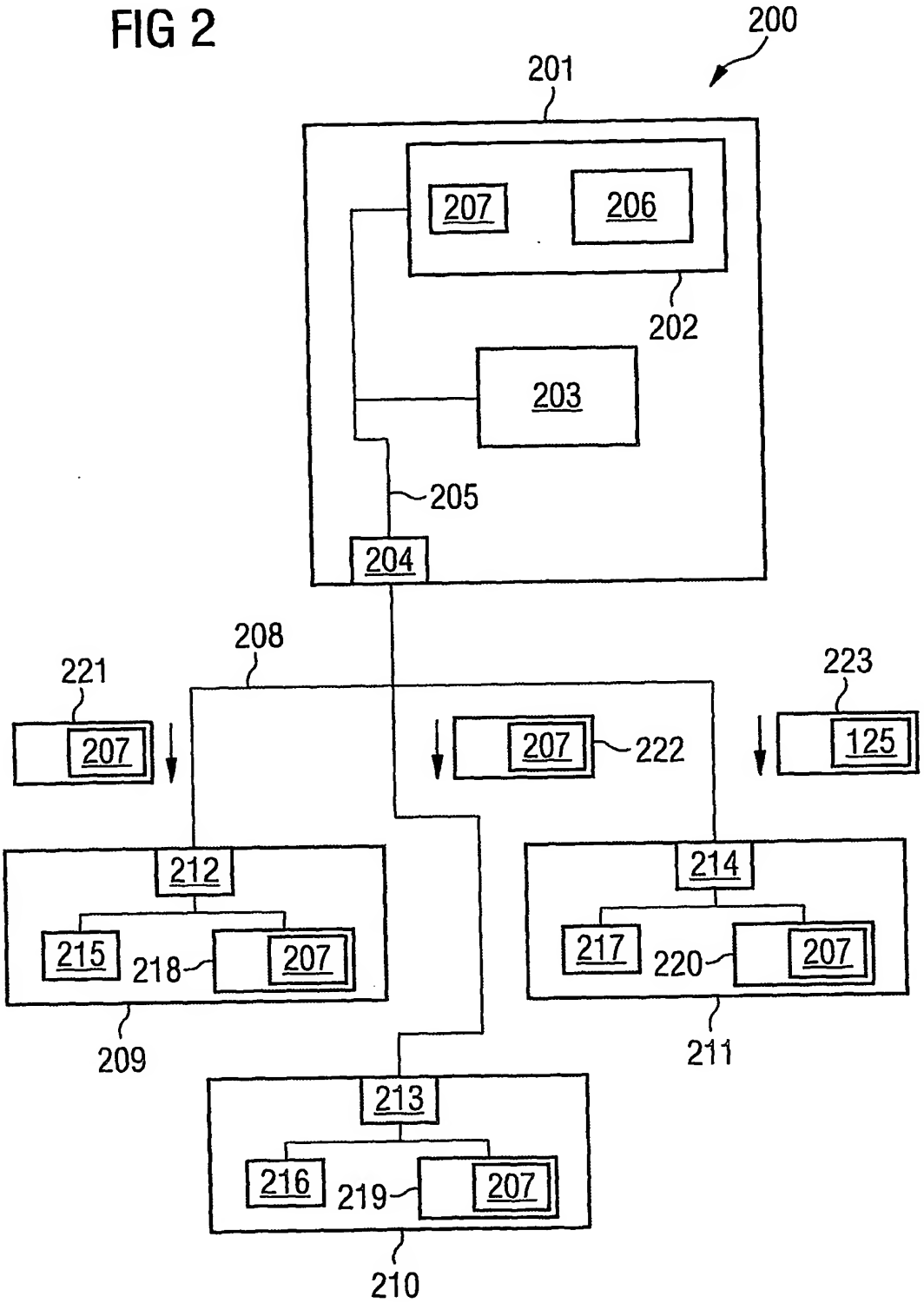

200
201
207
206
202
203
205
204
208
221
207
207
222
223
125
212
215
218
207
209
214
217
220
207
211
213
216
219
207
210

FIG 3

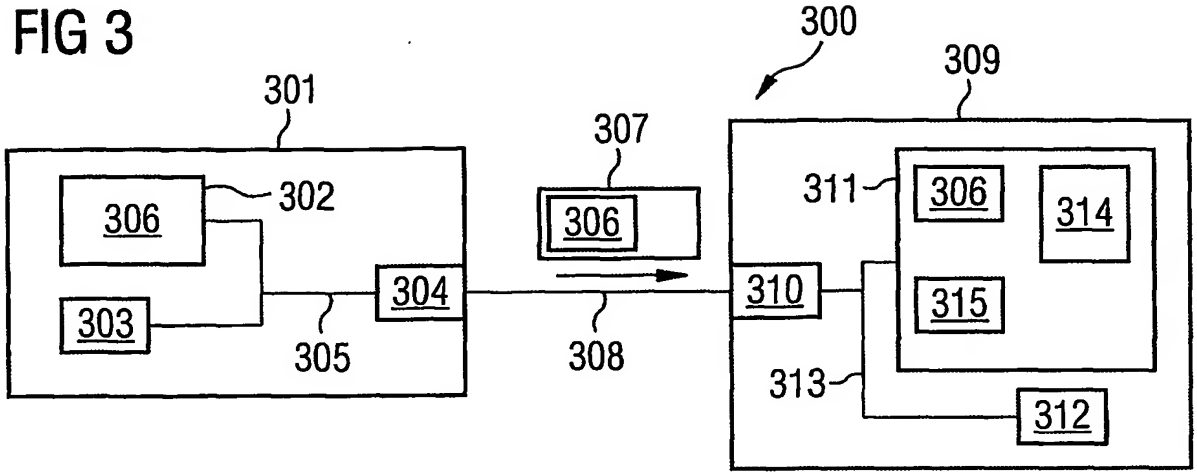

300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315

FIG 4

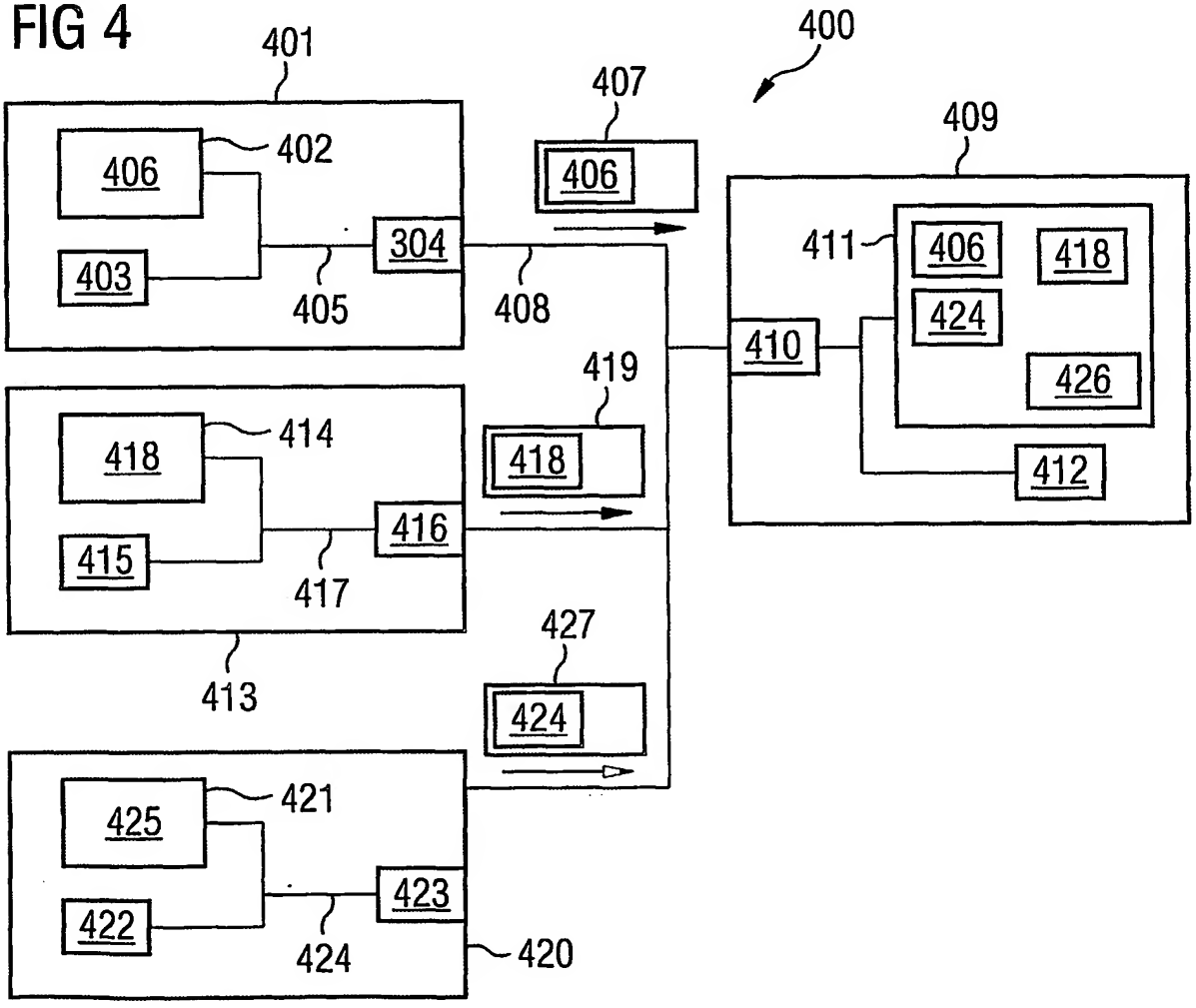
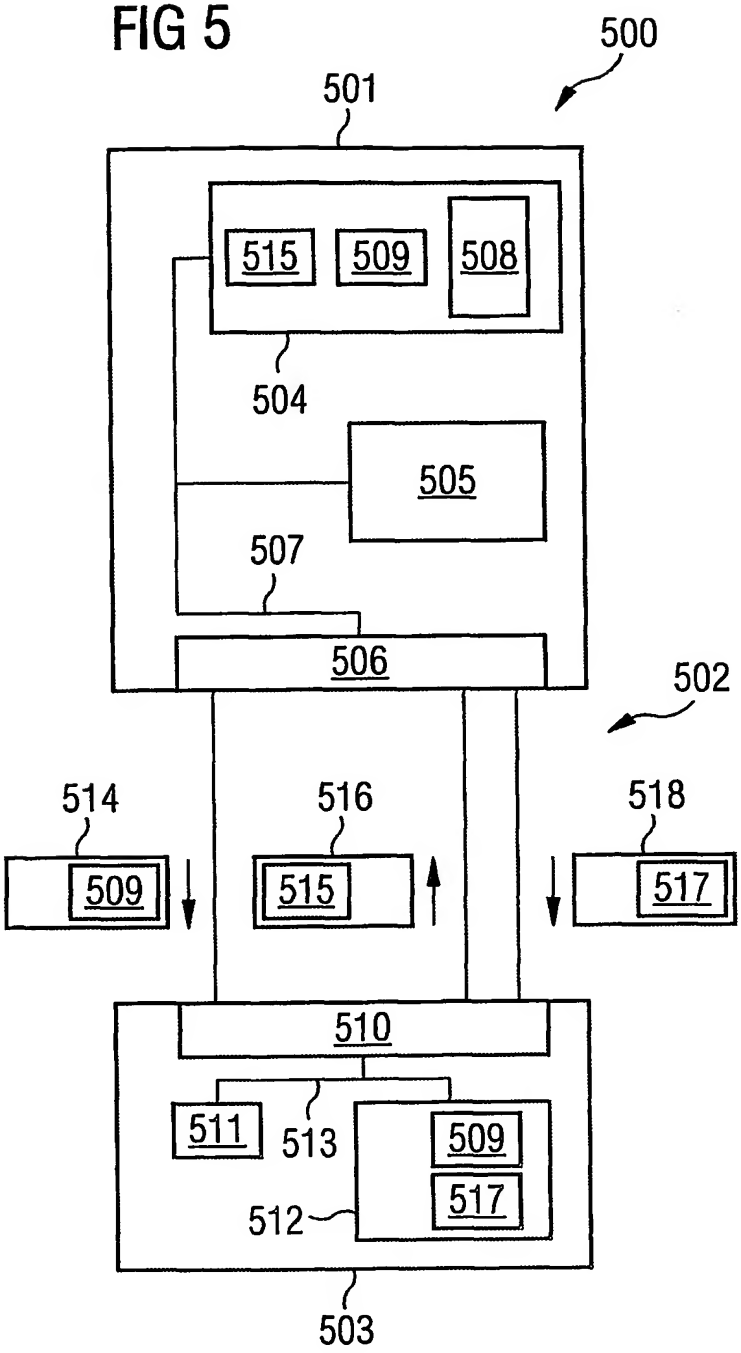

400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427

FIG 5


500
501
515
509
508
504
505
507
506
502
514
516
518
509
515
517
510
511
513
509
517
512
503